DOCUMENT RESUME

ED 367 699 TM 021 163

TITLE	The Trial State Assessment: Prospects and Realities. The Third Report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: 1992 Trial State Assessment.
INSTITUTION	National Academy of Education, Stanford, Calif.
PUB DATE	93
NOTE	221p.
AVAILABLE FROM	The National Academy of Education, Stanford University, School of Education CERAS-507, Stanford, CA 94305-3084.
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Statistical Data (110)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Academic Achievement; *Comparative Analysis; Educational Assessment; Elementary School Students; Elementary Secondary Education; Evaluation Methods; Federal Government; Financial Support; Government Role; Grade 4; Grade 8; Grade 12; *Mathematics Tests; National Surveys; *Reading Tests; Secondary School Students; State Programs; Testing Problems; *Testing Programs; *Test Use; Test Validity
IDENTIFIERS	National Academy of Education; *National Assessment of Educational Progress; Reform Efforts; *Trial State Assessment (NAEP)

ABSTRACT

The National Academy of Education Panel is providing the intellectual leadership and coordination for an independent evaluation of the 1990 and 1992 Trial State Assessments (TSA) of the National Assessment of Educational Progress (NAEP). These instruments provide state-by-state comparisons of educational achievement, which is the first such use of NAEP data. A portfolio is being constructed of studies of the TSA using a range of methodological approaches. In the 1992 TSA eighth-grade students were again tested in mathematics and fourth-grade students were assessed in mathematics and reading. Approximately 8,700 schools and 330,000 students participated in the 1992 TSA. Issues examined were the technical feasibility, content validity, utility of results, who should be assessed, and how results should be reported. The Panel believes that the full potential of the NAEP is not being realized because of funding constraints. Conducting both national and state assessments may be prohibitively expensive, but the Panel recommends that Congress provide adequate funding to allow the continuation of the national assessment while at the same time assessing at least one subject at three grade levels in state assessments. A number of recommendations are made for improvement of both types of assessment. Ten figures and 26 tables present findings. (Contains 63 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TM

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

L.S. SHULMAN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED 367 699

The Trial State Assessment: Prospects and Realities

The Third Report of
The National Academy of Education Panel
on the Evaluation of the NAEP Trial State Assessment:
1992 Trial State Assessment

Panel Chairmen
Robert Glaser, University of Pittsburgh
Robert Linn, University of Colorado

Project Director
George Bohrnstedt, American Institutes for Research

TM 021163

BEST COPY AVAILABLE

The Trial State Assessment: Prospects and Realities

Copyright © 1993 by The National Academy of Education
All rights reserved.
Printed on recycled paper in the United States of America
The National Academy of Education
Stanford University
School of Education CERAS-507
Stanford, CA 94305-3084
(415) 725-1003
Designed and printed by Armadillo Press, Palo Alto, CA

The Trial State Assessment: Prospects and Realities

The Third Report of
The National Academy of Education Panel
on the Evaluation of the NAEP Trial State Assessment:
1992 Trial State Assessment

Panel Chairmen
Robert Glaser, University of Pittsburgh
Robert Linn, University of Colorado

Project Director
George Bohrnstedt, American Institutes for Research

The National Academy of Education

The National Academy of Education is composed of scholars and education leaders who "promote scholarly inquiry and discussion concerning the ends and means of education, in all its forms, in the United States and abroad." Our current active membership is limited to 125 scholars. The heart of the Academy is found in the serious and vital exchanges that take place in our regular meetings and the special commissions we establish. Throughout our 28-year history, the Academy has been called upon by governmental and other agencies to conduct special studies and reviews in the public interest. The most comprehensive evaluation of the National Assessment of Educational Progress ever conducted was organized by the Academy and directed by Lamar Alexander and H. Thomas James in 1986. We not only served as fiscal agent and coordinator of the Alexander-James Panel, we also provided a thoroughly independent review of that endeavor. Most recently, we completed a study, funded by the Carnegie Corporation of New York, to recommend priorities for the funding of educational research entitled "Research and the Renewal of Education."

As the national trend toward education reform broadens and accelerates, we focus again on NAEP. This time we contribute the intellectual leadership and coordination for an independent evaluation of the 1990 and 1992 Trial State Assessments, which provide state-by-state comparisons of educational achievement, the first such use of NAEP data. Specifically, we are conducting a carefully constructed portfolio of studies using a range of methodological approaches. A panel of experts, selected by the Academy on the basis of their distinguished accomplishments, manages and coordinates the study portfolio. The Panel and the Academy have examined a range of major issues: the conduct and reporting of the assessment; the validity and reliability of the assessment data; the possibility of linking NAEP to state and international assessments; the relationship of NAEP to emerging national content standards; the inclusiveness of the assessment population; the continuation of a prohibition on reporting district or school-level data; and the viability of defining performance standards for student achievement.

The Panel submitted its first report, on the 1990 Trial State Assessments, in November of 1991. In July of 1993 it submitted a special study on the definition and use of performance standards for student achievement in the 1992 assessment. The current report, the Panel's third study, covers a wide range of issues pertaining to the 1992 Trial State Assessment. The Panel also plans a capstone report in which they will further explore several broad issues relevant to the evolution of the National Assessment of Educational Progress.

Carl F. Kaestle
President, The National Academy of Education

Table of Contents

Transmittal Letter xi

Acknowledgments xiii

Foreword xv

The Panel xvii

Executive Summary xix

1 *Introduction* 1

2 *The Conduct of the 1992 Assessment* 15

3 *The Content Validity of the Assessment* 49

4 *Reporting and Impact of State NAEP Results* 75

5 *Guiding Principles* 93

6 *Should State NAEP Be Continued?* 99

7 *Future of NAEP: Who Should Be Assessed?* 105

8 *How Should NAEP Results Be Reported?* 111

9 *Should the Prohibition Be Lifted?* 117

10 *How Should NAEP Frameworks Be Related to Emerging National Content Standards?* 121

11 *Can NAEP Be Linked Meaningfully to State and International Assessments?* 125

12 *How Can State Participation in State NAEP Be Maximized?* 131

13 *Prospects and Realities* 137

Appendix 145

Notes 189

Acronyms ..... Inside Back Cover

Detailed Table of Contents

Transmittal Letter xi
Acknowledgments xiii
Foreword xv
The Panel xvii

Executive Summary xix

1 Introduction 1

The Value of NAEP as an Indicator of Student Achievement 1
The Influence on NAEP of the Movement for National Education Standards 2
The 1990 Trial State Assessment and the Panel's Recommendations from its Evaluation 10
The 1992 Trial State Assessment 12
Important Policy Issues for the Future of NAEP 12
An Overview of This Report 13

2 The Conduct of the 1992 Assessment 15

The Trial State Assessment Sample 16
Sampling and Participation of Schools 17
Sampling and Participation of Students 21
Weighting 23
Excluded Students 24
The Administration of the 1992 TSA 35
Trial State Assessment and National NAEP Comparisons 36
Training and Monitoring of Assessment Administrators 36
Session Administration 37
State Testing Directors' Perspectives on the Implementation of the 1992 TSA 39
The Cost to the States for Participation 40
Summary and Recommendations 45

3 The Content Validity of the Assessment 49

Objectives and Frameworks for the 1992 Trial State Assessment 49
Mathematics Framework 50
Reading Framework 52
Implementation of the Frameworks—Items for the 1992 TSA 57
Mathematics Items 57
Reading Items 63
Directing the NAEP Development Effort 68
Summary and Recommendations 69

4 Reporting and Impact of State NAEP Results 75

Reporting the 1992 TSA in Mathematics 77
Reporting by Achievement Levels 79
State Response to Achievement Levels and Other New Report Features 80

Reporting the 1992 TSA in Reading 81
Delays in the Reporting Schedule 82
1992 Dissemination Activities 83
Impact of the TSA 83
Press Coverage 84
Evidence of Impact from the Reporting of the 1990 TSA 85
Summary and Recommendations 90

5 Guiding Principles 93

The National Indicator Principle 93
The Lead and Reflect Principle 94
The Policy Relevance Principle 95
The State Indicator Principle 95
The Inclusiveness Principle 96
The Relative Stability Principle 97
The International Indicator Principle 97
The Assessment Innovation Principle 97
Overview of the Remaining Chapters 98

6 Should State NAEP Be Continued? 99

The Technical Feasibility of State NAEP 99
The Sampling Design 99
Participation Rates 99
Comparing the State NAEP Sample to a National NAEP Composite Sample 101
The Effect of Monitoring Sessions 101
Exclusions Due to Limited English Proficiency and Individualized Education Plans 101
The Content Validity of the 1992 Mathematics and Reading Assessments 102
The Utility of State NAEP Results 103
Summary 104

7 Future of NAEP: Who Should Be Assessed? 105

Administering Special Versions of NAEP to Children with IEPs 105
Administering Special Versions of NAEP to Children with LEP 106
Administering NAEP to Out-of-School 17-Year-Olds 110

8 How Should NAEP Results Be Reported? 111

The Panel's Evaluation of the Achievement Levels 111
The Angoff Method Is Fundamentally Flawed 111
The Panel Endorses Standards-Based Reporting 112
New Achievement Levels Must Be Treated as a Research and Development Effort and Be Rigorously Evaluated 112
Achievement Levels Require that NAEP Content and Performance Standards Be Consistent with National Standards 113
International Comparisons Should Be Made 114
Move to Within-Grade Score Reporting 114

9 Should the Prohibition Be Lifted? 117

Continuing Calls for Lifting the Prohibition 117
Arguments in Favor of Lifting the Ban 118
Arguments Against Lifting the Prohibition 118

10 How Should NAEP Frameworks Be Related to Emerging National Content Standards? 121

Influence of the Standards Movement 121
NAEP Frameworks Must Lead and Reflect 122
NAEP Content Must Be Comprehensive 122
NAEP Frameworks Must Reflect National Performance and Content Standards 122
Standing Subject-Matter Panels Should Be Established 123

11 Can NAEP Be Linked Meaningfully to State and International Assessments? 125

Equating Tests with Common Frameworks with Relative Ease 125
The Challenge of Linking Assessments with Different Frameworks 125
The Research Base 126
Linking Standardized Mathematics Tests to the 1990 NAEP Mathematics Assessment 126
Examining the Content Congruence Between the Standardized Mathematics Tests and the 1990 NAEP Mathematics Assessment 128
Requirement of Common Frameworks for Linking 129
The Prospect of Linking NAEP to International Assessments 129

12 How Can State Participation in State NAEP Be Maximized? 131

The Burdens States Face in Participating in State NAEP 131
Incentives That Might Be Employed to Maximize Participation in State NAEP 132
Remove the Prohibition Against Reporting Data Below the State Level? 132
Allow States to Mix and Match Subjects and Grade Levels Assessed? Move to Annual NAEP Assessments? 132
Allow Reduced Student Samples for Small States? 133
Use Different Assessment Administrators? 134
Have the Federal Government Continue to Bear a Major Portion of the States' Costs for the Assessment? 134

13 Prospects and Realities 137

A Design for an Effective and Comprehensive NAEP 137
NAEP Would Be an Annual Assessment with Comprehensive Content and Frameworks 137
NAEP's Role as an Independent Monitor Would Be Protected 137
The Cost for State Participation Would Be Borne by the Federal Government 138
The Assessment Would Be More Inclusive 138
The Assessment Would Report What Students Should Know as Well as What They Actually Know 139
NAEP Would Continue to Be a Leader in Assessment Innovation 139

Difficult Choices Given Current Funding Realities 139
The Possibility of Aggregating State NAEP Samples to Obtain National Performance Estimates 140
Maintaining the Quality of Both State and National NAEP 141
A Concluding Thought 142

Appendix: Synopses of Background Studies 145

A Study of Eligibility Exclusions and Sampling: 1992 Trial State Assessment 146
A Study of Students Excluded from the 1992 National Assessment of Educational Progress Trial State Assessment 149
State and Local Costs of the 1992 Trial State Assessment 153
The Content and Curricular Validity of the 1992 NAEP Reading Framework 157
Evaluation of the 1992 Reading Framework for the National Assessment of Educational Progress 162
The Content and Curricular Validity of the 1992 NAEP in Mathematics 165
Content Validation of the 1992 NAEP in Reading: Classifying Items According to the Reading Framework 168
Impact of the 1990 NAEP Trial State Assessment Program: A Followup Study 172
Issues in the Development of Spanish-Language Versions of the National Assessment of Educational Progress 176
The Judged Congruence Between Various State Assessment Tests in Mathematics and the 1990 National Assessment of Educational Progress Item Pool for Grade-8 Mathematics 180
A Study of the Administration of the 1992 National Assessment of Educational Progress Trial State Assessment 183

Notes 189

Acronyms ..... Inside Back Cover

December 9, 1993

Emerson Elliott
Commissioner
National Center for Education Statistics
U.S. Department of Education
555 New Jersey Avenue, NW
Washington, DC 20208-5653

Dear Emerson:

On behalf of The National Academy of Education, I am pleased to transmit to you the third report of the Academy's Panel on the Evaluation of the NAEP Trial State Assessment. The report is entitled *The Trial State Assessment: Prospects and Realities*. This report focuses on the conduct, validity, and uses of the 1992 Trial State Assessments and then applies research findings to several major policy issues concerning the future of NAEP.

This report has been reviewed and approved by The National Academy of Education's Executive Council, acting as a Committee of Readers. Like the preceding reports of this panel, it should assist policymakers in reaching thoughtful, informed decisions about several important policy issues concerning the National Assessment of Educational Progress.

In addition to the complex and important policy issues that must be resolved in the short-run, the Panel and the Executive Council of The National Academy of Education are concerned about some broader issues that need to be considered in thinking about the longer-run future of NAEP. The Panel has broached some of these issues in the present or previous reports. In this report, for example, the Panel points out that some assessment items reflect older content while others reflect new content; aggregating them obscures that mixture, which is shifting across successive assessments. Achieving some short-term stability and comparability in NAEP while remaining sensitive to changing consensus on appropriate content standards will remain a major issue.

The Panel intends to consider other major issues in the capstone report scheduled for 1995. Among them are the following:

> NAEP measures certain forms of competence, but by no means all the forms of competence that educators consider important today. There is a danger of privileging certain kinds of learning and performance. We need more good research and thinking about competence and its assessment in content areas.
>
> Assessments based on multiple-choice tests do not yield information about the processes of problem-solving. NAEP is moving substantially away from multiple-choice questions; but further innovations in assessment methods will be needed to address current concerns about learning skills and the self-monitoring of one's performance.

The Panel and the Executive Council of The National Academy of Education therefore remind readers of this report not only that the immediate policy issues addressed here are complex and very consequential, but that the evolution of NAEP is taking place in a context of profound changes in the assessment field, and in the context of a reform movement that has made assessment a central but problematic tool. The Council is pleased that this distinguished panel is bringing to bear its expertise on these difficult questions, and we trust that all those interested in the future of NAEP will find their efforts illuminating and helpful.

Sincerely,

Carl Kaestle

Carl F. Kaestle
President, The National Academy of Education
William F. Vilas Professor of Educational Policy
Studies and History, University of Wisconsin

Acknowledgments

This mandated report presents the findings of the Panel of the National Academy of Education regarding the evaluation of the 1992 Trial State Assessment (TSA) of the National Assessment of Educational Progress (NAEP).

The Panel extends special thanks to the National Center for Education Statistics (NCES) for its oversight and support for the Panel's independent review. Throughout this project, Commissioner Emerson Elliott has provided valuable guidance. Steve Gorman, Gary Phillips, Larry Ogle, and Sharif Shakrani have been particularly helpful in providing advice and information to the Panel. Ed Mooney has continued to assist Panel staff with the administrative details of the project.

Without the generous support of the contractors for the 1992 NAEP assessment, Educational Testing Service (ETS), Westat, and National Computer Systems (NCS), this evaluation could not have been completed in a timely fashion. ETS staff have continued to provide a high level of prompt, professional support to the Panel and Panel staff. Among the many staff at ETS who responded to the Panel's requests for data and technical assistance were (in alphabetical order): Kent Ashworth, Jay Campbell, Jules Goodison, Eugene Johnson, Steve Lazar, Ina Mullis, John Olson, and Al Rogers. In addition, ETS has continued to keep the Panel informed about contractor meetings and data-collection activities. At Westat, Nancy Caldwell, Brice Hart, Keith Rust, and Dianne Walsh provided data files, details about the training and administration of the TSA, and information on the sampling design for the assessment. In addition, staff at NCS provided data for the Panel's study of the administration of the TSA.

Special thanks are also due to Jeremy Finn (State University of New York, Buffalo), who spent the summer at the American Institutes for Research and played a critical role in the formative stages of this report.

Thanks also to National Assessment Governing Board (NAGB) members Roy Truby, Dan Taylor, Mary Lyn Bourque, Ray Fields, and Mary Crovo for giving generously of their time in answering queries and in providing requested information.

The Panel would like to thank the state testing directors, state curriculum directors, and all other national, state, and local educators and policymakers who responded to its surveys on the administration of the 1992 TSA and the impact of reporting the 1992 TSA results.

The Panel expresses its gratitude to the distinguished researchers who carried out the commissioned investigations. Thanks are extended to Richard Jaeger (University of North Carolina, Greensboro), P. David Pearson, Lizanne DeStefano, Chip Bruce, and Jean Osborn (University of Illinois, Urbana-Champaign), Walter Secada (University of Wisconsin), Bruce Spencer (Northwestern University), Ed Silver and Pat Kenney (University of Pittsburgh), and Michelle Commeyras (University of Georgia). Summaries of the commissioned reports are included in the appendix of this volume.

Foreword

The National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment was formed to respond to the Congressional charge for an independent evaluation of the feasibility and technical adequacy of NAEP at the state level. In responding to this charge, the Panel has considered a wide range of factors that influence the quality of NAEP in its role as the primary means of indicating educational progress, both for the states participating in the Trial State Assessment and for the national assessment of achievement. Prominent considerations in the deliberations of the Panel include the content of the assessments, the inclusiveness of the population to be assessed, sample design, participation rates, reporting, and the uses made of results.

Three reports have been issued based on studies and analytical papers commissioned by the Panel. These papers and studies have informed the Panel's discussion and debate that has resulted in its conclusions and recommendations. The first report, *Assessing Student Achievement in the States*, issued in March 1992, covered the conduct of the 1990 assessment, its content validity, scoring and interpretation, and reporting the results to the public and press.

The second report, *Setting Performance Standards for Student Achievement*, issued in September 1993 at the request of the Commissioner of the National Center for Education Statistics, studied the achievement levels used for interpreting NAEP results. This report examined the process used for setting achievement levels, the validity and reasonableness of the achievement levels in reading and mathematics for 1992, and the relationship of NAEP to national education standards.

The current, third report comes at a time when decisions and trade-offs must be made about the extent and cost feasibility of both state and national NAEP. Compromises must be made that do not damage the accuracy and usefulness of assessments conducted nationwide at the state level. This report, *The Trial State Assessment: Prospects and Realities*, examines state NAEP 2 years later—the 1992 state assessment. It looks again at issues of sampling and administration, content validity, and reporting impact. These findings, together with earlier findings of the Panel's studies and recommendations made in the earlier reports, provide the basis for discussion of policy issues and guiding principles that comprises a major emphasis of this third report. The intention of the Panel is to present the issues and questions involved in a way that is most informative to technical, policy, and fiscal decisions that need to be made by Congress, NCES, NAGB, and NAEP contractors.

It is to be emphasized, once again, that the Panel is aware that its work is taking place in a period of heightened activity and change with respect to the assessment and interpretation of student achievement. Numerous innovations are being implemented at the national, state, and local levels, and the Panel views NAEP as an extremely valuable, independent monitor of the status and trends in change in student achievement as our nation proceeds toward education reform. Continued evaluation and study of the nation's key indicator of educational progress is an ongoing endeavor that can supply information to technical, professional, and policymaking communities and can contribute to making the improvement of our educational systems one of the nation's outstanding accomplishments. As this occurs, change is in the wind; testing

technology is undergoing critique and renewal involving new standards of validity and reliability. New methodology and models are being developed to assess competence in thinking, problem solving, and the skills of learning. NAEP must lead this change while it also engages the major policy issues considered in this report.

Robert Glaser, Chairman
Robert Linn, Co-Chairman

December 14, 1993

The National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment Project

Robert Glaser, Chairman
Director, Learning Research and Development Center (LRDC)
and National Research Center on Student Learning
University of Pittsburgh

Robert Linn, Chairman
Co-director, National Center for Research
on Evaluation, Standards, and Student Testing
University of Colorado at Boulder

Gordon Ambach
Executive Director
Council of Chief State School Officers

Isabel Beck
Professor of Education
Senior Scientist, LRDC
University of Pittsburgh

Lloyd Bond
Professor, Educational Research Methodology
University of North Carolina at Greensboro

Ann Brown
Professor of Education in Math, Science, and Technology
Evelyn Lois Corey Fellow in Instructional Science
University of California at Berkeley

Iris Carl
National Council of Teachers of Mathematics (NCTM)

Alonzo Crim
Professor of Education
Spelman College

Robert M. Groves
Professor of Sociology and Associate Director
Joint Program in Survey Methodology
University of Michigan

Edward Haertel
Professor of Education
Stanford University

Lyle Jones
Professor of Psychology and Director
L.L. Thurstone Psychometric Laboratory
University of North Carolina at Chapel Hill

Edward Roeber
Director of Student Assessment Programs
Council of Chief State School Officers

Albert Shanker
President
American Federation of Teachers

Lorrie Shepard
Professor of Education
University of Colorado at Boulder

Project Staff
American Institutes for Research

George Bohrnstedt, Project Director
Don McLaughlin, Associate Project Director
Jay Chambers
Shannon Daugherty
Phyllis DuBois
Dey Ehrlich
Jeremy Finn
Elizabeth Hartka
Susan Kleimann
Elise McCandless
Cathy O'Donnell
Frances Stancavage
Jean Wolman

Learning Research and Development Center

Elizabeth S. Rangel
Cindy Yockel

Executive Summary

The Trial State Assessment: Prospects and Realities

The National Assessment of Educational Progress (NAEP) has been the "nation's report card" for almost 25 years, reporting on the academic proficiency of the country's 4th, 8th, and 12th graders. NAEP is an unparalleled source of information about U.S. students' academic achievement in important subject areas, currently including reading, writing, mathematics, science, history, and geography. The NAEP sampling design enables it to accurately describe the academic proficiency of U.S. students in grades 4, 8, and 12. Moreover, it provides information about specific subpopulations of students by gender, race or ethnicity, region of the country, and size or type of community. Compared to alternative data sources such as college admissions tests or commercially published achievement tests, NAEP has provided more broadly useful performance data because of its national probability sample and high participation rates. NAEP offers broad coverage of curriculum in core subject areas, and it is designed to permit meaningful trend measurement so that changes over time can be identified.

In 1988, as part of Public Law 100-297, Congress mandated trial state versions of NAEP for 1990 and 1992 to provide states with information about their students' achievement and, secondarily, to allow them to compare their students' performance with that of students in other states. Implementation was on a trial basis so that feasibility and utility could be determined before resources were committed to an ongoing program of state-by-state assessments. Through Congressional authorization, the National Academy of Education (NAE) was commissioned to conduct an evaluation of the NAEP program at the state level. NAE established an independent panel to undertake this work.

The NAE Panel's Previous Two Reports

This is the NAE Panel's third report on the evaluation of the Trial State Assessment (TSA). The first report, which evaluated the 1990 TSA, concluded that Congress should continue state NAEP but recommended additional trials before making state NAEP permanent.[1] The Panel noted that at the conclusion of the 1992 trial, only two subjects at grade 4 (mathematics and reading) and one subject at grade 8 (mathematics) would have been evaluated. The Panel addressed important policy issues in its first report, making a recommendation against lifting the prohibition that would allow NAEP data to be reported below the state level, as well as recommendations for the establishment of standing content-area committees to ensure continuity within and across assessments in NAEP content areas, the inclusion of private school students and out-of-school 17-year-olds in state NAEP, and the use of NAEP scale scores and anchor points to report results regardless of whether achievement levels were reported.

[1] The National Academy of Education, *Assessing Student Achievement in the States* (Stanford, CA: Author, 1992).

The National Assessment Governing Board (NAGB), the body that Congress charged to oversee and make policy decisions affecting the national assessment, made a decision to report the national and state 1990 mathematics results using achievement levels on a trial basis. Achievement levels are a type of standards-based reporting, indicating what children should be able to do at described levels of achievement. Although many NAEP constituencies, including the states, favor standards-based reporting, the implementation of the NAGB decision to report NAEP results by achievement levels proved to be controversial. A General Accounting Office (GAO) report found the approach used to set the 1990 and 1992 levels "procedurally flawed" and found the results based on the levels to be of "doubtful validity."[2] Because two other evaluations also found technical flaws with the achievement levels set in 1990,[3] the National Center for Education Statistics (NCES) asked this Panel to conduct its own independent evaluation of the 1992 achievement levels. The evaluation culminated in the Panel's second report, *Setting Performance Standards for Student Achievement,* which concurred with earlier evaluations that found the achievement levels to be flawed and recommended that the Governing Board suspend their use.[4] The Panel also made several specific recommendations about approaches to use in attempting to establish performance standards on NAEP for the future.

The 1992 Trial State Assessment

The 1992 TSA, the subject of this report, was both a continuation and an extension of the 1990 assessment. In 1992, eighth-grade students were again assessed in mathematics; in addition, fourth-grade students were assessed in mathematics and reading. Because two grades and two subjects were assessed, the magnitude of the assessment task was substantially greater in 1992. The number of schools participating in state NAEP increased from approximately 3,600 to about 8,700, and the number of students assessed increased from approximately 100,000 to about 330,000. The number of states that chose to participate also increased from 37 to 41.[5]

[2] U.S. General Accounting Office, *Educational Achievement Standards: NAGB's Approach Yields Misleading Interpretations*. Report No. GAO/PEMD-93-12 (Washington, D.C.: Government Printing Office, June 1993).

[3] R.L. Linn, D.M. Koretz, E.L. Baker, and L. Burstein, *The Validity and Credibility of the Achievement Levels for the 1990 National Assessment of Educational Progress in Mathematics* (Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing, January 1991) and D. Stufflebeam, R.M. Jaeger, and M. Scriven, *Summative Evaluation of the National Assessment Governing Board's Inaugural 1990-91 Effort to Set Achievement Levels on the National Assessment of Educational Progress* (Washington, D.C.: National Assessment Governing Board, August 1991).

[4] The National Academy of Education, *Setting Performance Standards for Student Achievement* (Stanford, CA: Author, 1993).

[5] In addition, two territories and the District of Columbia participated in each of the TSAs.

Important Policy Issues for the Future of NAEP

In conducting the evaluation of the 1992 TSA, the Panel focused much of its attention on technical issues such as the adequacy of the sampling frame, participation rates, the impact of random monitoring on the quality of the assessment sessions, and the validity of the assessment content frameworks. However, important *policy* issues also arose in evaluating the quality and integrity of the TSA. The most important policy questions addressed in this report are

- Should a continuing state NAEP be authorized? Have the 1990 and 1992 TSAs shown that state NAEP is technically feasible, content valid, and useful?
- Who should be assessed? In particular, should special versions of NAEP be administered to those limited-English-proficient (LEP) students and those students with individualized education plans (IEP) who cannot participate in the regular assessment? Should dropouts be included?
- How should NAEP results be reported, given the problems that have been found with the current achievement levels?
- Should the prohibition against reporting NAEP data below the state level be lifted? Should large districts that can meet the sampling requirements for participation be allowed to obtain district data? Should the design of NAEP be changed to support reporting by school?
- What should the relationship be between the content NAEP assesses and the voluntary national content standards being developed? How can continuity between NAEP content, classroom practice, and national standards be ensured?
- What are the prospects for linking NAEP to state and international assessments? What are the requirements and potential problems?
- How can NAEP ensure the fullest participation of states and schools? What burdens result from participation? What incentives might be offered?
- Are there cost savings or other efficiencies to be gained by building state NAEP as the foundation of national NAEP? Is this feasible?
- What would be the ideal NAEP assessment program, given no budget limitations? What choices are ahead, given funding realities?

In addition to these questions, this report addresses a number of design changes that have been proposed for NAEP, such as assessments in more subject areas, annual

assessments, and flexibility to allow states to choose the subjects and grade levels in which their students will be assessed.

The Panel's Guiding Principles

To help frame analyses and weigh the various policy issues and options, the Panel draws on its previous two reports by articulating a set of general principles, which are summarized below.

The National Indicator Principle. NAEP should continue to be the key indicator of what the nation's students know and can do, providing trend data on student academic performance in key subject areas.

The Lead and Reflect Principle. NAEP frameworks and content must be comprehensive, reflecting the most forward-looking pedagogical approaches, at the same time that they reflect the best of current practice.

The Policy Relevance Principle. NAEP must collect timely and relevant data for policymakers but not have its integrity corrupted by political pressures.

The State Indicator Principle. While NAEP's main role is that of the key nationwide indicator of educational progress, an additional needed role is to provide information to the states on the educational progress of their students.

The Inclusiveness Principle. To the degree technically, ethically, and financially possible, NAEP should assess an inclusive sample of all children whose ages would ordinarily place them in the 4th, 8th, or 12th grade.

The Relative Stability Principle. NAEP frameworks should remain unchanged for a reasonable period of time. The same framework must be in place long enough to establish meaningful trend lines, but not so long as to fail to accommodate the natural evolution that occurs in subject-matter fields.

The International Indicator Principle. Given the need for the United States to learn more effectively from educational practices in other nations and to compare practices and results, NAEP should, to the extent possible, be linked with major, relevant, large-scale international assessments.

The Assessment Innovation Principle. NAEP must continue to be the nation's leader in assessment innovation.

The Panel's Findings and Recommendations

The Panel has developed a set of recommendations for the Commissioner of Education Statistics to forward to Congress and the states. These recommendations are based on findings from the studies and papers the Panel commissioned to address many of the technical and policy issues outlined above. At this time, the Panel offers these recommendations in the hope that they provide useful information as Congress carries out its deliberations on the reauthorization of NAEP.

Should State NAEP Be Continued?

The most important question addressed by the Panel was whether to recommend the continuation of state NAEP. In formulating its recommendation, the Panel examined the technical feasibility of state NAEP, the content validity of the 1992 mathematics and reading assessments, and the usefulness of NAEP results for the states.

Technical Feasibility

In judging the technical feasibility and soundness of the 1992 TSA, the Panel considered the adequacy of the sampling design; school and student participation rates; session administration, including differences in the performance of students in the TSA versus the national assessment, and in monitored versus unmonitored sessions of the TSA. State-by-state patterns of student exclusions based on individualized education plans (i.e., disabilities) or limited English proficiency were also explored.

Sampling design. The sampling design was judged to be of high quality, and the assessment itself was carefully executed by the contractor. However, the sampling size requirements placed a burden on some small states.

Participation rates. The percentage of initially sampled schools that agree to participate is called the *initial participation rate*. When schools refuse to participate, substitute schools also chosen by the contractor are then contacted by the state coordinator for participation in the study. A state's *final participation rate* is the number of schools that were initially contacted and eventually participate plus those that are contacted as substitute schools and also participate.

The average initial school participation rates in the 1992 TSA were adequate compared to those achieved for national NAEP in recent years. However, compared to the 1990 TSA, there was increased *variability* between states in the initial school participation rates for 1992, and more states had before-substitution rates below 85 percent. In 1990, there were only two states with initial school participation rates below 85 percent; in the 1992 TSA, one-third of the states were in this category. The lowest 1992 participation rates were for Maine, which recruited only 62 percent of its originally sampled schools in grade 8 and 58 percent in grade 4.

BEST COPY AVAILABLE

For 1992, NAEP reports flagged results for states that had initial weighted school participation rates below 85 percent *and* final weighted school participation rates below 90 percent, as well as for states that had initial weighted school participation rates below 70 percent. In addition, states could be flagged if the nonparticipating schools were concentrated within a particular class of schools, or if the weighted *student* participation rate fell below 85 percent.[6] These flags appeared only in the appendix of the 1992 mathematics reports but were in the main text of the 1992 reading reports. For 1994, NAGB has passed a resolution indicating that states failing the overall participation-rate guidelines should be reported in a special section, separate from those states meeting or exceeding the participation criteria. In the Panel's view, the criteria used for noting states with low participation rates are too lenient. Further, the Panel agrees that the results for these states should be reported in a separate section, not simply flagged. **Because there is a substantial likelihood of nonresponse bias when initially selected schools fail to participate, the Panel recommends that the results from states with initial school participation rates below 85 percent be reported separately from the results of states with initial rates of at least 85 percent, regardless of the state's final school participation rates.**

The potential seriousness of the low initial participation rates in some states is underscored by the Panel's finding that higher performance on the assessment was associated with lower participation rates. This finding was replicated in both the fourth- and eighth-grade samples, suggesting that the result is not due to chance. Furthermore, the Panel's analyses suggest that the finding is not due to one or two aberrant states. *Thus, there is a concern that states with low initial participation rates might have inflated results on NAEP, and the Panel finds some of the states' initial participation rates too low for accurate reporting of their 1992 TSA results.* The Panel will continue to examine the process or processes that might account for the negative correlation between initial participation rates and NAEP performance. In the meantime, the Panel is concerned that the credibility of the TSA could be jeopardized if this problem is not corrected. Therefore, **the Panel recommends that the contractor take steps to ensure that uniform procedures are followed by the state coordinators in recruiting schools to participate in state NAEP.**

Comparing the state NAEP sample to a national NAEP composite sample. In its evaluation of the 1990 TSA, the Panel found a small but reliable difference between the mathematics performance of eighth graders in the TSA sample and a specially constructed matched subset of students in national NAEP. For the 1992 administration, no reliable differences between the three TSA samples and corresponding composite national samples were observed. Because the 1990 result suggested that there may have been differences in student motivation (e.g., students taking the TSA might be more highly motivated to do well) or differences in administration procedures (e.g., the use of local administrators in the TSA versus the use of administrators hired by the contractor in the national administration), the results from the 1992 administration were reassuring.

The effect of monitoring sessions. In order to ensure the quality of the administration of the assessment at the local level, 50 percent of the sessions, chosen randomly, were monitored by quality control monitors. The results of the monitoring

[6] For grade 4, five states were flagged because they failed to satisfy the guideline for overall participation rates, and one additional state was flagged because of under-representation of a class of schools. An overlapping group of six states was flagged at grade 8. No states were flagged for violating the student participation guideline.

indicate that the administration of the assessment was adequate in the vast majority of sessions. Importantly, no differences in overall student performance were found between the monitored and unmonitored sessions in either the 1990 or 1992 TSA.

IEP and LEP exclusions. Overall, about 2 to 3 percent of students sampled were excluded because of limited English proficiency; exclusions for LEP students ranged from less than 1 percent in some states to nearly 11 percent in California in the fourth-grade reading assessment. Overall, about 5 to 6 percent of the students sampled were excluded because of an individualized education plan. IEP exclusion rates varied from 2 to 7 percent across the states. Substantial consistency between the 1990 and 1992 TSA exclusion rates was observed, suggesting that the same states that were high (or low) in exclusion rates in 1990 tended to be high (or low) in 1992 as well. Between-state variations in LEP exclusions are to be expected, given the demographic differences among the states, but additional research is needed to understand why some states exclude more IEP students than others when there is no a priori reason to expect such differences.

In summary, based on its evaluation of sampling, participation, and administration procedures in the 1990 and 1992 trials, the Panel concludes that state NAEP is technically feasible. However, more research is needed before a firm conclusion can be drawn. The Panel has identified major concerns related to participation rates. The negative association between initial school participation rates and average state assessment scores is potentially a serious matter; therefore, NCES and its contractor should immediately undertake research that examines this finding. In addition, the current guideline for flagging the results of schools with low initial school participation rates is too lenient and should be made more stringent.

The Content Validity of the 1992 Mathematics and Reading Assessments

Mathematics. In the Panel's judgment, the 1990/1992 Mathematics Framework and item pool have shown progress in moving toward an assessment that incorporates the standards developed by the National Council of Teachers of Mathematics (NCTM) but still reflect too much of what is currently being taught and do not extend far enough to include content implied by the NCTM Standards.[7] Despite incremental improvements, the fact that fewer than 10 percent of the 1992 items were judged as exemplars of the NCTM Standards provides evidence that the NAEP mathematics assessment needs to change considerably before it fully meets the Panel's lead and reflect principle. The 1994 Mathematics Framework is expected to provide substantial improvement in this respect.[8]

Reading. The 1992 reading assessment, based on a newly developed framework, is a significant improvement over previous NAEP reading assessments. The assessment makes excellent use of authentic passages and open-ended items. Although the reading community, unlike the mathematics community, has yet to coalesce around a

[7] National Council of Teachers of Mathematics, *Curriculum and Evaluation Standards for School Mathematics* (Reston, VA: Author, 1989).

[8] Although a 1994 framework was developed for mathematics, there will not be a 1994 mathematics assessment. The next mathematics assessment is now expected to occur in 1996.

single statement of standards, experts from the reading field generally viewed the framework as a reasonable balance between what would be ideal and what is currently being taught at the classroom level. However, as was also found to be true in mathematics, additional challenging items need to be added at the upper levels of the NAEP reading scale.

The Utility of State NAEP Results

The Panel's studies of the 1990 and 1992 mathematics assessments show that the states strongly support a state NAEP. Perhaps the most compelling evidence for this is the fact that the number of states participating in the TSA increased from 1990 to 1992, and this number appears likely to increase again for 1994.[9]

Evidence from surveys and case studies conducted by the Panel in 1992 further suggests that the 1990 TSA in mathematics has had a positive impact on curriculum and assessment in many of the states that chose to participate. Because so many influences are operating at once, it is difficult to attribute changes solely, or even primarily, to the TSA. However, 46 percent of the respondents from states that were involved in curricular or instructional reform in mathematics *and* that had participated in the 1990 TSA indicated that NAEP and the TSA influenced their reform efforts. Respondents credited the TSA with adding impetus to already planned or desired changes, accelerating discussions of state goals, and increasing pressure to align with national standards.

The Panel's Recommendation on Continuing State NAEP

Based on its evaluation of the 1990 and 1992 trials, the Panel recommends that Congress authorize a continuation of state NAEP. However, there is a need for ongoing evaluation. Certain questions related to full-scale implementation of state NAEP have not been answered by the first two trials and will not be answered by the limited, one-grade, one-subject trial scheduled for 1994. Both the technical feasibility and the perceived burden to the states of conducting a state NAEP at three grade levels in multiple subjects remain to be evaluated. Also, there has been no 12th-grade trial in any subject area, and it may be more difficult to secure participation from 12th-graders. **For these reasons, the Panel recommends that Congress mandate an ongoing evaluation of state NAEP with regular feedback to Congress. Because many of the factors affecting the quality and feasibility of state NAEP are the same as those affecting national NAEP, the Panel notes that evaluation of the two programs could be integrated for maximum efficiency and effectiveness.**

[9] At the time of this writing, 46 states and territories have signed up for the 1994 TSA.

Who Should Be Assessed?

The Panel's first report emphasized the principle of inclusiveness, and this principle assumes even more importance with the increases in the numbers of students with disabilities who have an individualized education plan and of students who have limited English proficiency. Some of the racial and minority groups for which NAEP provides separate estimates of achievement are disproportionately represented in these populations, further adding to the importance of inclusiveness. The Panel's studies indicate that overall about 5-6 percent of the students in fourth and eighth grade were excluded from the 1992 TSA on the basis of IEPs and about 2-3 percent on the basis of LEP, figures that underestimate because they exclude students who are not enrolled in regular public schools.

Students with IEPs. The principle of inclusiveness would support including as many IEP students as possible in the NAEP assessment, and it seems probable that adapted, parallel assessments could be constructed and used effectively with many of these students. However, such a plan would be quite costly. Labor-intensive individual assessments would probably be required for many learning disabled students; only partial testing would be possible for other students; and no type of assessment would be possible for a small number of cases in which disabilities are especially severe.

Prior to taking any steps toward assessing additional IEP students, the Panel believes that more research is needed to determine the severity of disabilities among those students enrolled in regular public schools who are presently excluded from NAEP. To do so, the Panel proposes to undertake two studies as part of the 1994 assessment. The Panel will report on the results of its research and will make a recommendation based on that research.

Students with LEP. Only about 3 percent of fourth graders and 2 percent of eighth graders were excluded from the 1992 TSA for reasons of limited English proficiency, but the rate varied significantly from one state to another, ranging from less than 1 percent in Alabama to about 5 percent of the eighth graders and 10-11 percent of the fourth graders in California. While these differences are not surprising, given the ethnic compositions of the states, the differences have an impact on the ability to make valid cross-state comparisons.

Spanish is by far the most common language spoken in the United States (after English), accounting for more than half (17,339,000 or 54.4 percent) of those who speak a language other than English at home and more than 10 times the number who speak the next language most commonly spoken. The Panel commissioned a paper to examine the pros and cons of developing a Spanish-language version of NAEP, and strong arguments were presented on both sides.[10] The Panel realizes that establishing a non-English version of NAEP would be a huge step with financial and political ramifications above and beyond the technical issues. For these reasons, *the Panel supports the development of a research agenda to examine the feasibility of developing a Spanish-language version of NAEP.*

[10] W.G. Secada, "Issues in the Development of Spanish-Language Versions of the National Assessment of Educational Progress," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

The Panel recommends that as part of the 1996 national assessment, NAGB authorize a set of pilot studies in locales where a significant proportion of students speak Spanish as their first language (e.g., Arizona, California, Florida, Illinois, New Mexico, New York). The pilot studies should include an experimental design that allows NCES to determine whether NAEP results vary as a function of administering a Spanish-language version of NAEP for Spanish-speaking students.

Out-of-school 17-year-olds. In its evaluation of the 1990 TSA, the Panel recommended that state NAEP reflect the achievement of all 4th-, 8th-, and 12th-grade students, not just those enrolled, pointing out the significant proportion of young people who drop out of school. The Panel recommended that beginning with the 1994 assessment, a national sample of out-of-school 17-year-olds be assessed, a recommendation that has not been implemented. On the basis of its inclusiveness principle, the Panel continues to believe that it is important to learn what the nation's out-of-school 17-year-olds know and can do. However, the Panel also recognizes that identifying and testing this sample could be prohibitively expensive. **If Congress fails to provide sufficient funding for NAEP to assess out-of-school 17-year-olds, the Panel recommends that future adult literacy surveys oversample 16-to-18-year-olds, stratifying on whether they are in or out of school, in order to provide the nation with a better picture of the literacy of the nation's out-of-school youth. Further, selected blocks of NAEP items should be included on the adult literacy survey to increase comparability between the two assessments. Finally, given that the proportions of out-of-school 17-year-olds vary across time and across the states, the Panel recommends that all 12th-grade state NAEP results released in the future also include information on the percentages of 17-year-olds who have not completed high school and that these percentages be reported alongside each state's results.**

How Should NAEP Results Be Reported Given the Problems with the Current Achievement Levels?

Because of the salience of the standards issue and controversy surrounding evaluations of the achievement levels as a primary reporting device for the 1992 TSA results, the Panel was asked to expand its evaluation of the 1992 TSA to include an extensive study of the 1992 achievement levels. The Panel has issued a separate report on its evaluation of the 1992 achievement levels and the issues surrounding their use.[11] The Panel's strong criticism of the achievement levels was based on the way in which they were developed, not on achievement levels per se.

The Panel recommended that NAGB discontinue the use of the Angoff method to set achievement levels, discontinue reporting using the flawed 1992 levels, and not reinstate achievement levels as the primary reporting mechanism until the following conditions are met: (1) a new, more defensible way of reporting the achievement levels can be developed, (2) the achievement levels have been empirically evaluated prior to using them as the primary NAEP reporting mechanism, and (3) assessment content is better articulated with emerging national content and performance standards.

[11] The National Academy of Education, *Setting Performance Standards*.

In addition, the Panel suggested alternate approaches that could be used. The Panel also recommends that whenever possible, international comparisons should be provided to help evaluate the performance of U.S. students.

Should the Prohibition Against Reporting NAEP Data Below the State Level Be Lifted?

The Panel included in its 1990 TSA evaluation a recommendation that the prohibition against reporting NAEP data below the state level be retained.[12] In this report, the issues for and against lifting the ban have been re-examined in light of calls by NAGB and the U.S. Department of Education for lifting the ban. Because of the danger of corrupting national NAEP, the country's key indicator of educational progress, the Panel reaffirms its position against removing the ban.

The Panel recommends that Congress continue the prohibition against reporting NAEP data below the state level. If Congress weighs and reads the evidence presented in this report and decides to lift the ban, the Panel recommends that (1) the ban against reporting data at the *school level* remain in place, and (2) that district-level NAEP be implemented *on a trial basis* subject to the following four conditions. First, the same objectives, booklets, and manipulatives as are used in national NAEP would have to be used in district-level NAEP. Second, the sampling plan used would have to meet appropriate statistical requirements. Third, administration procedures for district-level NAEP would have to be identical to those used in state NAEP. Fourth, districts that participate should be required to help fund and participate in a Congressionally mandated independent evaluation.

How Should NAEP Frameworks Be Related to Emerging National Content Standards?

To remain credible, NAEP frameworks and items must be informed by the emerging national content standards and assessment practices. Changes in the frameworks and items have occurred since NAEP's inception, and alignment with national content standards has already begun. As these changes take place, it is important that NAEP frameworks be made comprehensive, reflecting current research and theory at the same time that they reflect the best of current classroom practice.

The issues surrounding assessment require a coordinated continuity of effort. In its previous two reports, the Panel recommended the establishment of standing subject-matter committees, and the Panel reiterates and elaborates that recommendation here, believing that these committees can be the mechanism for linking the comprehensive content envisioned in *Goals 2000* to NAEP, while also ensuring that NAEP continues to reflect current classroom practice. **The Panel recommends that NAGB establish**

[12] Panel member Gordon Ambach supported lifting the prohibition at that time and continues to hold that position. He is on record elsewhere as recommending the lifting of the prohibition at the district level where the size of the enrollment allows sampling as used at the state level. Panel member Edward Roeber dissents as well.

standing subject-matter panels. In addition to subject-matter experts and researchers, membership in the panels should include classroom teachers, members of the public with expertise in the subject area, and representatives of NAGB. These standing panels should participate in the development of new frameworks, and when relevant, should consider explicitly how to incorporate certified national standards. Given the likely difference between the content based on nationally certified standards and that based on current classroom practice, the panels also should advise NCES whether to combine assessment results from the two content domains into a single composite score or whether to report them as two separate scores. Finally, the panels should provide continuity to the assessment by being involved in all aspects of the process, including formulating the framework and objectives; reviewing items, item-scoring rubrics, and reporting formats; and helping to achieve agreement on narrative descriptions of performance standards and representative illustrative tasks. This would help to ensure that there is both logical and statistical correspondence between content standards, performance standards, and illustrative tasks.

Can NAEP Be Linked Meaningfully to State and International Assessments?

Interest in how U.S. students compare with students in other countries has fueled an interest in linking state and national assessments with international assessments. States are keenly interested in linking their state assessments not only to NAEP but also to international assessments such as the upcoming Third International Mathematics and Science Study (TIMSS). The ability to make *meaningful* links between assessments presents considerable, although not insurmountable, challenges.

Tests and assessments designed to a common framework can be linked with relative ease; linking assessments with different frameworks is more problematic. Successful linking requires that the constructs assessed, the content frameworks, the difficulty distributions of the items, and the reliabilities of the two assessments be the same. The Panel monitored a study conducted by the Technical Review Panel that examined the feasibility of linking statewide tests to NAEP when all of the technical requirements of equating could *not* be met;[13] the Panel also commissioned a companion study to better understand the influence of content on the dependability of linking results.[14] The empirical results of these two papers support the previous findings and suggest that *state assessments using standardized achievement tests may not be able to establish a trustworthy link between their assessments and NAEP. Furthermore, by implication, states will be better able to equate their state assessment results to NAEP if their assessments weigh content areas in a manner consistent with NAEP frameworks.* Furthermore, while more research is needed, *it seems likely that a common framework, and items that are sampled from that framework, are necessary conditions for linking state assessments to NAEP.*

[13] R.L. Linn and V.L. Kiplinger, "Linking Statewide Tests to the National Assessment of Educational Progress: Stability of Results" (Boulder, CO: Center for Research on Evaluation, Standards, and Student Testing, 1993).

[14] L. Bond and R.M. Jaeger, "The Judged Congruence Between Various State Assessment Tests and the 1990 National Assessment of Educational Progress Item Pool for Grade-8 Mathematics," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

Linking NAEP to international assessments (e.g., TIMSS) raises similar technical issues. **Based on its research, the Panel recommends that TIMSS either use a framework similar to the NAEP framework or that a subset of the TIMSS assessment be mapped onto the NAEP framework so that scores with comparable weighting of content from the TIMSS can be equated to NAEP.**

How Can State Participation in State NAEP Be Maximized?

The Panel identified a number of obstacles to state participation in the 1992 state NAEP. While the number of participating states increased from 37 in 1990 to 41 in 1992, this is still substantially fewer than all 50. The proportion of initially sampled schools refusing to participate in state NAEP increased in 1992, and refusals rose to unacceptable levels in several states.

States that agree to participate in state NAEP must meet the contractor's sampling requirements for district and school participation. This task is particularly difficult in small states which must rely on the same schools year after year and, in some cases, must test the same students in more than one subject. Furthermore, beginning in 1994, the states must also secure the participation of private schools in state NAEP. The Panel considered various incentives that might be employed to maximize participation in state NAEP.

One possibility is to remove the prohibition against reporting the data below the state level. As detailed above, the Panel objects to removing the prohibition because of a concern that doing so could corrupt the national assessment.

Another way to reduce the burden on states would be to allow them to choose the subjects and grade levels to be assessed—sometimes referred to as the "mix-and-match" option. In addition, the burden could be reduced by moving to annual assessments instead of assessments every other year, with fewer subjects per year. While there is merit to the option of mixing and matching subjects and grades, part of the public appeal for state NAEP is based on the reporting of data for *all* participating states. With mixing and matching, state-by-state results could be so badly compromised as to render state-by-state reporting totally meaningless, severely eroding public support. Regarding annual assessments, **while it is not in the Panel's purview to decide on which subjects should be assessed and when, the Panel recommends allowing NAEP assessments to be administered each year as a way to effectively distribute the administrative work over the years and reduce the burden on states in any given year.**

Another proposal would be to allow small states to use smaller samples of students in sampled schools. **Assuming that stable statistical estimates can be based on samples of reduced size, the Panel recommends that small states be allowed to use reduced samples of students when participating in state NAEP.**

Having the federal government bear the states' costs for the assessment would be a large incentive for participation. However, the U.S. Department of Education's draft legislation could do just the opposite. In particular, the legislation would increase the Secretary of Education's authority to require states to pay for some or all of the cost of scoring, analyzing, and reporting state NAEP data, costs currently paid by the federal

government. **The Panel recommends that the federal government not move additional costs for state NAEP to the states. While the motivation for the proposed legislation may be to protect state NAEP, implementation of the option to bill the states for NAEP-related costs might actually result in fewer states participating in state NAEP.**

Prospects and Realities

The Panel believes the full potential of NAEP is not being realized because of the funding constraints under which NAEP will be operating at least through the end of the 1994 assessment. It has been suggested both to and within the Panel that mounting two assessments—national and state NAEP—is not cost efficient. A seemingly obvious solution would be to aggregate the results from state NAEP in such a way as to create national NAEP estimates with supplemental samples drawn from those states not participating in state NAEP.

The first step in deciding whether this suggestion is feasible rests on examination of the purposes and uses of national and state-level data. NAEP is the single best indicator of what the nation's children know and can do, and it is imperative that its integrity be preserved. There may be content areas of interest to Congress that may not be of high priority for the states. Furthermore, even within a given subject area, national NAEP is not merely a single sample that parallels that of the TSA. National NAEP includes a large number of bridge studies used for charting trends, as well as other special studies conducted on a periodic or as-needed basis. It would not be cost effective or, in some cases, even feasible to construct the samples for these ongoing and special studies by aggregating the state NAEP samples.

For all these reasons, it seems unlikely that national NAEP could or should be totally replaced. However, cost efficiencies might be gained using an aggregation of state NAEP samples to obtain national population estimates in *selected grades and content areas*.

Aggregating state NAEP samples to obtain national estimates has other problems, however. First, the level of voluntary participation by states, year after year in assessments involving multiple content areas and grades, remains unknown. The fewer the states that participate in state NAEP, the more inefficient the idea of deriving national performance estimates from an aggregation of state NAEP samples becomes. Second, states would have to make firm commitments to participate in NAEP several months in advance. A backup process would need to be in place in case states decided to drop out at a late date or in case they failed to achieve minimum school participation requirements. Third, the cost of state NAEP is very high relative to the cost of national NAEP. For 1994, the cost of state NAEP for one subject at one grade level is estimated to be \$9-10 million.[15] Adding either another subject at the same grade level or a second grade level adds an additional \$4-5 million to the cost of the assessment. The cost of national NAEP for three grades and three subjects for 1994 is estimated at \$10-12 million. The cost of aggregating the 1994 state NAEP into a national NAEP (assuming three subjects were assessed at three grade levels) would be between \$41 and \$50 million, which *excludes* the cost of drawing supplemental

[15] The cost information in this section comes from discussion with NCES staff. The cost figures are for data collection and analysis only. The development costs, which run from \$3-4 million per subject, exist whether there is a state NAEP or not.

samples for nonparticipating states.[16] Hence, the cost of a national NAEP aggregated from a state NAEP consisting of three subjects at three grade levels would be more than four times the cost of the current national NAEP and would be reasonable only if state NAEP were fully funded.

In summary, it seems unlikely that national NAEP could ever be fully replaced by aggregating state samples. However, if state NAEP were conducted on an annual basis with a high level of state participation, cost savings *might* be realized and the burden on states reduced by estimating performance on national NAEP through an aggregation of state NAEP samples. Therefore, **the Panel recommends that NCES examine whether cost efficiencies could be realized by a partial substitution in which aggregation of state NAEP samples would be used to make national estimates of student performance, supplemented as necessary with samples from nonparticipating states. The aggregation of state results for national estimates in a given grade and content area would not rule out the possibility of national NAEP being carried out in other content areas or at other grade levels.**

Maintaining the Quality of Both National and State NAEP

The Panel has not addressed the issue of how much NAEP would cost if its full potential were to be realized. Merely to *maintain* the current quality of NAEP will require additional funding. The 1994 NAEP had been planned for five subjects at the national level and two at the state level—a reasonable plan in the Panel's view. However, the funding appropriated for 1994 NAEP will allow for only three subjects at the national level and one at the state level, and in the latter case, at only one grade level. Therefore, the full potential for both national and state NAEP is far from being realized. **The Panel recommends that Congress provide more adequate funding for future NAEP assessments—funding that will allow national NAEP to continue to be a high quality, comprehensive indicator of the educational achievement of all the nation's children while at the same time providing for an annual state NAEP in *at least* one subject at three grade levels. The funding required to achieve these goals is relatively modest when compared to the benefit that the nation and states will realize from the assessments.**

[16] This figure is derived by assuming that the first subject at one grade level costs $9-10 million and adding to it $4-5 million for *each* of the additional eight subject-by-grade assessments.

1 Introduction

Since first administered to the nation's students in 1969, the National Assessment of Educational Progress (NAEP) has been an important indicator of educational achievement; its data have highlighted trends in the progress, decline, or steady state of student performance over the years. NAEP has both reflected and, at times, promoted changes in subject-matter content and assessment methodology, always with the aim of improving its effectiveness as "the nation's report card." In reviewing the history of NAEP and evaluating its use at the state level, this Panel finds that the assessment program is at a crossroads. Because of tight funding for NAEP, difficult choices will have to be made. The choices that lie ahead have critical implications for the path that NAEP will follow during the remainder of this century and beyond.

One strong influence affecting current demands on the NAEP program is the emerging national interest in establishing standards and criteria to evaluate educational performance at the international, national, and state levels. The President, the nation's governors, representatives of business, professional educators' organizations, the research community, and the general public have all contributed to this movement. For example, President Clinton has proposed putting the six education goals formulated at the 1989 Education Summit into law as part of his administration's *Goals 2000: Educate America Act*. This legislation, if passed, would add impetus to the development of voluntary academic standards and assessments and would pave the way for identifying the conditions that ensure opportunities for all students to meet high performance standards.

Support for the establishment of national standards and for assessments of progress in meeting those standards is widespread. The assessment most frequently suggested to measure progress at the state and national levels is NAEP.

The Value of NAEP as an Indicator of Student Achievement

NAEP is an unparalleled source of information about U.S. students' academic achievement in reading, writing, mathematics, science, history, and geography. NAEP's sampling design enables it to accurately describe the academic proficiency of U.S. students in grades 4, 8, and 12 without having to assess every student. Moreover, it provides information about specific subpopulations of students by gender, race or ethnicity, region of the country, and size or type of community. Compared to alternative data sources such as college admissions tests or commercially published achievement tests, NAEP has provided more broadly useful performance data because of its national probability sample and high participation rates.

NAEP offers broad coverage of curriculum in core subject areas, and it is designed to permit meaningful trend measurement so that changes over time can be identified. For example, NAEP trend data indicate that student proficiency in science declined in

the 1970s but increased significantly between 1982 and 1990.[1] NAEP results for subpopulations revealed that the science proficiency of 9-year-old black students was greater in 1990 than in 1970 and that, for ages 9 and 13, the gap between the science proficiency of black and white students decreased from 1970 to 1990.[2]

NAEP assessments also include questionnaires to gather data on background variables related to student, teacher, and school characteristics, as well as curriculum and instruction. For example, NAEP data indicate that during the 1980s students in all three age groups (9, 13, and 17 years) increased the amount of time they spent watching television. In 1982, 55 percent of students 13 years old reported that they watched 3 or more hours of television per day; in 1990, that percentage had increased to 70.[3] Such data shed light on some of the factors that can affect achievement.

In 1990, the Trial State Assessments (TSA) broadened the scope of the NAEP program to provide results that are representative at the state level as well as at the national level. For example, policymakers, educators, and parents in Colorado can now compare their students' academic proficiency to that of students in each of the neighboring states of New Mexico, Arizona, Utah, Wyoming, and Nebraska, as well to the academic proficiency of students in the United States as a whole.

The high quality of NAEP's instruments and assessment methodology, along with its history of providing useful data, make it the logical instrument for measuring U.S. students' progress toward standards at the state and national levels. The Panel has recommended previously that "NAEP should be preserved as the nation's key independent indicator of how students are performing academically,"[4] and it continues to support that recommendation vigorously. The primary purpose of this report is to evaluate the feasibility of state NAEP; however, with the increase in demands and expectations concerning the NAEP program overall, the issues involving state NAEP have become more complex. Those issues provide the context for this report of the 1992 Trial State Assessment.

The Influence on NAEP of the Movement for National Education Standards

Figure 1.1 summarizes the history of NAEP and the concurrent moves toward national standards. This figure shows that the first voluntary national content standards, developed by the National Council of Teachers of Mathematics (NCTM), appeared in

[1] I.V.S. Mullis, J.A. Dossey, M.A. Foertsch, L.R. Jones, and C.A. Gentile, *Trends in Academic Progress: Achievement of U.S. Students in Science, 1969-70 to 1990; Mathematics, 1973 to 1990; Reading, 1971 to 1990; Writing, 1984 to 1990* (Washington, D.C.: National Center for Education Statistics, November 1991), 34.

[2] Ibid., 35.

[3] Ibid., 100.

[4] The National Academy of Education, *Assessing Student Achievement in the States* (Stanford, CA: Author, 1992), 67.

final form in 1989.[5] The next year, in 1990, both national and trial state assessments in NAEP mathematics were administered. Despite the brief time lapse between these two events, the NCTM Standards made a marked contribution to the content framework for the 1990 mathematics assessment. The NCTM Standards were also used as the basis for adding new item formats to the 1992 mathematics assessment, and the 1994 NAEP Mathematics Framework was changed to align still more closely with the NCTM Standards. Similarly, the 1994 History and Geography Frameworks were influenced by the work being done to establish national standards in both those areas.

Figure 1.1. ***Important events in the development and use of the National Assessment of Educational Progress***

1964 Initial assessment planning and development activities begin for the National Assessment of Educational Progress (NAEP), with support from the Carnegie Corporation and the Ford Foundation. NAEP has these features: NAEP is an "assessment," not a test; no total scores are reported, only reports of achievement on individual items; a consensus process is used to formulate objectives and create exercises; and reporting is by four geographic regions only (not by states, districts, or schools).

1969 The National Assessment of Educational Progress programs begin by testing the science, writing, and citizenship proficiency of 17-year-olds in both public and private schools, out-of-school 17-year-olds, and adults 26-35 years old. NAEP is administered by the Education Commission of the States (ECS).

1970 NAEP assesses science, writing, and citizenship proficiency of students ages 9, 13, and 17, out-of-school 17-year-olds, and adults 26-35 years old.

1971 NAEP assesses reading and literature proficiency of students ages 9, 13, and 17, out-of-school 17-year-olds, and adults 26-35 years old.

NAEP is now financed solely by federal funding.

1972 NAEP assesses music and social studies proficiency of students ages 9, 13, and 17, out-of-school 17-year-olds, and adults 26-35 years old.

1973 NAEP assesses mathematics and science proficiency of students ages 9, 13, and 17, out-of-school 17-year-olds, and adults 26-35 years old.

[5] National Council of Teachers of Mathematics, *Curriculum and Evaluation Standards for School Mathematics* (Reston, VA: Author, 1989).

Figure 1.1. ***Important events in the development and use of the National Assessment of Educational Progress (continued)***

1974	NAEP assesses career and occupational development of students ages 9, 13, and 17, out-of-school 17-year-olds, and adults 26-35 years old and assesses writing proficiency of students ages 9, 13, and 17 and out-of-school 17-year-olds.
1975	NAEP assesses reading and art proficiency of students ages 9, 13, and 17 and out-of-school 17-year-olds and basic skills proficiency of students age 17 and out-of-school 17-year-olds. Secondary-use data files are produced, enabling outside researchers to use the NAEP database.
1976	NAEP assesses citizenship/social studies proficiency of students ages 9, 13, and 17 and out-of-school 17-year-olds and assesses consumer mathematics proficiency of students ages 13 and 17.
1977	NAEP assesses science proficiency of students ages 9, 13, and 17, basic life skills proficiency of students age 17, and science, reading, and health proficiency of adults ages 26-35 years old.
1978	NAEP assesses mathematics proficiency of students ages 9, 13, and 17 and consumer skills proficiency of students age 17.
1979	NAEP assesses writing, art, and music proficiency of students ages 9, 13, and 17.
1980	NAEP assesses reading/literature proficiency of students ages 9, 13, and 17 and out-of-school 17-year-olds and art proficiency of students age 13.
1982	NAEP assesses mathematics and citizenship/social studies proficiency of students ages 9, 13, and 17 and science proficiency in a special study of students ages 9, 13, and 17.
1983	Educational Testing Service (ETS) assumes responsibility for administering NAEP and implements a planned schedule of biennial assessment, assessment of reading every biennium, coverage of four subject-matter fields as a target for each assessment, increased sampling efficiencies through balanced incomplete block (BIB) spiraling, sampling by grade level as well as by age, random sampling within schools to reduce sampling errors, use of behavioral anchor reporting in results, scaling by item response theory (IRT) technology, and simplified reports, targeted to lay audiences.

Figure 1.1. ***Important events in the development and use of the National Assessment of Educational Progress (continued)***

1983 *(cont.)*	Publication of *A Nation at Risk* increases the calls for education reform, which are fueled by declines in NAEP science and mathematics achievement and other education indicators.[6]
1984	NAEP assesses reading and writing proficiency of students ages 9, 13, and 17; grade samples of students (in grades 4, 8, and 11) are added.
	NAEP adopts a scaling procedure that produces a standard score scale ranging theoretically from 0 to 500, with a median of 250 and a standard deviation of 50 (for grades 4, 8, and 12 combined) in the base year. The scale is held constant after the base year so that year-to-year changes can be seen. Descriptions of the types of things that students know and can do at each major score interval are developed for anchor points on the scale (200, 250, 300, and 350).
	Secretary of Education Terrell H. Bell develops the *Wall Chart* to provide state-by-state data on performance and suggests that ACT and SAT scores are the best achievement data available for this purpose.
1985	The Council of Chief State School Officers (CCSSO) votes to support the expansion of NAEP as one mechanism for collecting state-level data.
	Florida, Tennessee, and Virginia become the first states to participate in state-test-to-NAEP comparisons for 11th-grade reading through the ETS-coordinated Southern Regional Education Board (SREB) project. Wyoming and New York also obtain NAEP assessment results in reading for their 11th-grade students.
	For the 1984 assessment, ETS produces its first public-use data files, with a new, easier-to-use format.
	NAEP assesses literacy in a sample of young adults, ages 21-25.

[6] National Commission on Excellence in Education, *A Nation at Risk* (Washington, D.C.: U.S. Department of Education, 1983).

Figure 1.1. ***Important events in the development and use of the National Assessment of Educational Progress (continued)***

1986 NAEP assesses mathematics, reading, science, and computer-competence proficiency of students ages 9, 13, and 17, grades 3, 7, and 11 and U.S. history and literature proficiency of students age 17, grade 11.

In 13 states, NAEP assessments that provide state-level data are conducted: 8 states use the SREB model first used in 1985; 1 state uses a variation of that model; 2 states contract with a third party to administer NAEP to all students in grades 3, 7, and 11; 1 state uses a 20 percent sample of students; and 1 state has a NAEP sample large enough to be representative of the students in the state.

Reading scores for 9- and 17-year-olds show substantial decline (for 13-year-olds, a slight increase), but mathematics scores do not decline. The anomaly leads to a detailed scrutiny of the reading assessment.

Secretary of Education William Bennett forms a blue-ribbon panel to make recommendations for the improvement of educational assessment.

1987 The blue-ribbon panel's findings are issued in the Alexander-James report titled *The Nation's Report Card,* which recommends that NAEP be revised and expanded to provide state-by-state comparisons.[7] A commentary by the National Academy of Education (NAE) accompanies the report and considers a range of issues important to NAEP, including a list of "necessary conditions" for the future success of NAEP.

A Technical Review Panel (TRP) is formed by the National Center for Education Statistics (NCES) to examine the accuracy of NAEP data and the potential problems in expanding NAEP to make state-by-state comparisons.

1988 NAEP assesses reading, writing, civics, and U.S. history proficiency of students ages 9, 13, and 17, grades 4, 8, and 12, and, with limited samples, document literacy of students ages 13 and 17, grades 8 and 12; geography proficiency of students age 17, grade 12; and mathematics and science proficiency of students ages 9, 13, and 17 and grade 11.

[7] L. Alexander and H.T. James, *The Nation's Report Card: Improving the Assessment of Student Achievement* (Washington, D.C.: The National Academy of Education, 1987).

Figure 1.1. ***Important events in the development and use of the National Assessment of Educational Progress (continued)***

1988 *(cont).*	The National Assessment of Educational Progress Act (P.L. 100-297) is enacted, placing responsibility for NAEP at NCES, authorizing trial state assessments for 1990 and 1992, providing for an independent evaluation of this state assessment pilot program, and establishing a National Assessment Governing Board (NAGB) to set policy direction for NAEP.
1989	President Bush and the nation's governors hold an Education Summit in Charlottesville, Virginia, and set goals for the year 2000, including a goal for America's students to demonstrate competency in English, mathematics, science, history, and geography at grades 4, 8, and 12. This and other reform efforts increase pressure for a national examination to provide evidence of what students know and can do.
	NAGB decides to "develop grade-level achievement goals that are 'under the umbrella' of broad national goals."[8]
	The National Council of Teachers of Mathematics (NCTM) publishes *Curriculum and Evaluation Standards for School Mathematics,* providing a broad outline of what mathematics should be taught at each grade level.[9]
	The NAE Panel is created to provide an independent evaluation of the extension of NAEP to the state level.
1990	The National Education Goals Panel (NEGP) is formed to measure progress toward education goals for the year 2000 and to issue reports on that progress at least annually. NEGP plans to use NAEP results and to use and emphasize achievement levels.
	NAEP assesses mathematics, reading, and science proficiency of students ages 9, 13, and 17, grades 4, 8, and 12, in about 1,500 public and private schools. Long term trends are assessed in mathematics, reading, writing, and science.
	The first NAEP Trial State Assessment (TSA) is conducted in mathematics, with the participation of 100,000 eighth graders in public schools in 37 states, the District of Columbia, and 2 territories.

[8] R. Truby, "The Future of NAEP as the Nation's Report Card" (Washington, D.C.: National Assessment Governing Board, November 29, 1989), 10.

[9] National Council of Teachers of Mathematics, op. cit.

Figure 1.1. ***Important events in the development and use of the National Assessment of Educational Progress (continued)***

1991	The 1990 NAEP mathematics results for the nation and the states are reported in June using the traditional anchor levels.[10] In September, mathematics results are reported by NAGB using the newly developed achievement levels on a trial basis.[11]
	NEGP releases its first "report card," also using a version of the new achievement levels.[12]
	Legislation is passed that establishes a National Council for Education Standards and Testing (NCEST) to provide advice to governors, Congress, and the President on the desirability and feasibility of education standards and testing.
1992	The NAE Panel issues its report evaluating the 1990 NAEP Trial State Assessment.[13]
	NCEST issues a report endorsing the development of education standards.[14]
	Nationally, NAEP assesses mathematics, reading, and writing at grades 4, 8, and 12. Long term trends are assessed in mathematics, reading, writing, and science.
	The second NAEP TSA is conducted in fourth-grade and eighth-grade mathematics and fourth-grade reading, with the participation of approximately 330,000 fourth and eighth graders in public schools in 41 states, the District of Columbia, and 2 territories.

[10] I.V.S. Mullis, J.A. Dossey, E.H. Owen, and G.W. Phillips, *The STATE of Mathematics Achievement* (Washington, D.C.: National Center for Education Statistics, June 1991).

[11] M.L. Bourque and H.H. Garrison, *The LEVELS of Mathematics Achievement* (Washington, D.C.: National Assessment Governing Board, September 1991).

[12] National Education Goals Panel, *The National Education Goals Report: Building a Nation of Learners* (Washington, D.C.: Author, 1991).

[13] The National Academy of Education, op. cit.

[14] National Council on Education Standards and Testing, *Raising Standards for American Education* (Washington, D.C.: Author, January 24, 1992).

Figure 1.1. ***Important events in the development and use of the National Assessment of Educational Progress (continued)***

1993	State and national results from the 1992 mathematics and reading assessments are released, using achievement levels as the primary reporting mechanism.[15,16]
	The U.S. General Accounting Office issues its report on NAEP achievement levels.[17]
	The NAE Panel issues its evaluation of reporting by the use of achievement levels.[18]
	President Clinton transmits to Congress *Goals 2000: Educate America Act,* which includes plans to develop voluntary academic standards and assessments for all students through the National Education Standards and Improvement Council (NESIC).
	Funding is authorized to assess reading, U.S. history, and world geography at grades 4, 8, and 12 in the 1994 national component of NAEP and reading at grade 4 in the 1994 Trial State Assessment of NAEP.
	The American Association for the Advancement of Science (AAAS) issues "guidelines" for elementary and secondary science instruction with descriptions of desired student outcomes.

Another very significant impact that the standards movement has had on NAEP was reflected in the 1989 decision of the National Assessment Governing Board (NAGB) to report assessment results in terms of what students *should* be able to do at described levels of achievement. While the concept of standards-based reporting has been well received, the implementation of the NAGB decision to report NAEP results by achievement levels proved to be controversial. A General Accounting Office (GAO) report found the approach used to set the 1990 and 1992 levels to be "procedurally flawed" and found the results based on the levels to be of "doubtful validity."[19] Because two other evaluations also showed technical flaws with the levels set in

[15] I.V.S. Mullis, J.A. Dossey, E.H. Owen, and G.W. Phillips, *NAEP 1992 Mathematics Report Card for the Nation and the States* (Washington, D.C.: National Center for Education Statistics, April 1993). In addition, separate reports were issued for each of the participating states and jurisdictions.

[16] I.V.S. Mullis, J.R. Campbell, and A.E. Farstrup, *NAEP 1992 Reading Report Card for the Nation and the States* (Washington, D.C.: National Center for Education Statistics, September 1993). In addition, separate reports were issued for each of the participating states and jurisdictions.

[17] U.S. General Accounting Office, *Educational Achievement Standards: NAGB's Approach Yields Misleading Interpretations.* Report No. GAO/PEMD-93-12 (Washington, D.C.: Government Printing Office, June 1993).

[18] The National Academy of Education, *Setting Performance Standards for Student Achievement* (Stanford, CA: Author, 1993).

[19] U.S. General Accounting Office, op. cit.

1990,[20] the National Center for Education Statistics (NCES) asked this Panel to conduct its own independent evaluation of the 1992 achievement levels. The evaluation culminated in the Panel's second report, *Setting Performance Standards for Student Achievement,* which concurred with earlier evaluations that found the achievement levels to be seriously flawed.[21] The results and recommendations from the Panel's report on achievement levels are summarized in chapter 8 of this report.

The 1990 Trial State Assessment and the Panel's Recommendations from Its Evaluation

In its original authorizations, NAEP was restricted to reporting summary results only for the nation as a whole and for large geographic regions. This changed in 1988, however, when NAEP was authorized to collect and report state-level data on a trial basis in 1990 and 1992. State participation was to be voluntary. In all, 37 states, the District of Columbia, and 2 territories agreed to participate in the 1990 state assessment in eighth-grade mathematics.

Under a grant from NCES, the National Academy of Education (NAE) established a Panel to conduct an evaluation of the 1990 and 1992 TSAs. NAE's evaluation of the 1990 TSA was released in February 1992, when the 1992 trial was already well underway. The evaluation found that the 1990 trial assessment was carried out successfully; however, because only a single trial had been studied, some questions about the technical quality and usefulness of state assessments could not be evaluated. The Panel therefore recommended that the trial be continued in expanded form (three subjects at three grade levels) as part of the 1994 assessment.[22]

The Panel also made a number of specific recommendations for improving state NAEP. As summarized below, these recommendations have informed the planning process for 1994.

- The Panel's recommendation that the trial be continued for one more round of assessments was adopted, although funds available for the 1994 assessment were not sufficient for the full expansion that the Panel had recommended.
- The Panel's recommendation to include private school students in state NAEP was implemented by Congress.

[20] R.L. Linn, D.M. Koretz, E.L. Baker, and L. Burstein, *The Validity and Credibility of the Achievement Levels for the 1990 National Assessment of Educational Progress in Mathematics* (Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing, January 1991) and D. Stufflebeam, R.M. Jaeger, and M. Scriven, *Summative Evaluation of the National Assessment Governing Board's Inaugural 1990-91 Effort to Set Achievement Levels on the National Assessment of Educational Progress* (Washington, D.C.: National Assessment Governing Board, August 1991).

[21] The National Academy of Education, *Setting Performance Standards.*

[22] The National Academy of Education, *Assessing Student Achievement.*

- The Panel's recommendation to continue random monitoring of schools participating in state NAEP was adopted.

- The Panel's suggestion that NCES develop new reporting devices and spread the release of results across the year was adopted, beginning with the reporting of the 1992 results.

- The Panel's suggestion that NAEP content be comprehensive, reflecting the most forward-looking and modern approaches while at the same time reflecting current classroom instruction, closely parallels a recommendation adopted by NAGB in March 1993.[23]

- The Panel's recommendation to evaluate the exclusion of children with disabilities from the NAEP assessment on a state-by-state basis will be carried out by the Panel itself, concurrent with the 1994 NAEP assessment. NAGB has begun consideration of this important issue as well.

- The Panel's recommendation against lifting the current prohibition against the reporting of data below the state level is under review. At this time, the prohibition remains in place for 1994; however, the Department of Education's draft plans for the reauthorization of NAEP would allow districts and schools to report NAEP results below the state level in the future.

- The Panel's recommendation that NAGB establish subject-matter committees to ensure continuity from the development of the frameworks through the various stages of item specification, item writing, item scoring, and the reporting of the results was not adopted.

- The Panel's recommendation that 17-year-olds no longer enrolled in school be assessed was not adopted.

- The Panel's suggestion that NAEP data be released using NAEP mean scale scores and their associated anchor levels as the primary reporting mechanism was not adopted.[24]

[23] The first sentence of the pertinent NAGB recommendation reads: "The content of the National Assessment should reflect both current and evolving instructional practice." National Assessment Governing Board, *Positions on the Future of the National Assessment of Educational Progress, 1993* (Washington, D.C.: Author, March 6, 1993), 4.

[24] All of these recommendations are in the NAE report *Setting Performance Standards*.

The 1992 Trial State Assessment

The 1992 TSA, which is the subject of this report, was both a continuation and an extension of the 1990 assessment. In 1992, eighth-grade students were again assessed in mathematics, and fourth-grade students were assessed in mathematics and reading. Because two grades and two subjects were assessed, the magnitude of the assessment task was substantially greater in 1992. The number of schools participating in state NAEP increased from approximately 3,600 to about 8,700, and the number of students assessed increased from approximately 100,000 to about 330,000. The number of states that chose to participate also increased from 37 to 41.[25]

Important Policy Issues for the Future of NAEP

Although, as noted, the press toward education standards is having considerable influence on NAEP, it is by no means the only issue of importance for the future of NAEP. Many questions arise in evaluating the quality and integrity of a national assessment program. NAEP-related policy issues discussed in this report include

- Should a continuing state NAEP be authorized? Have the 1990 and 1992 TSAs shown that state NAEP is technically feasible, content valid, and useful?
- Who should be assessed? In particular, should special versions of NAEP be administered to those limited English proficient (LEP) students and those students with individualized education plans (IEP) who cannot participate in the regular assessment? Should dropouts be included?
- How should NAEP results be reported given the problems that have been found with the current achievement levels?
- Should the prohibition against reporting NAEP data below the state level be lifted? Should large districts that can meet the sampling requirements for participation be allowed to obtain district data? Should the design of NAEP be changed to support reporting by school?
- What should be the relationship between the content NAEP assesses and the voluntary national content standards being developed? How can continuity between NAEP content, classroom practice, and national standards be ensured?
- What are the prospects for linking NAEP to state and international assessments? What are the requirements and potential problems?

[25] In addition, two territories and the District of Columbia participated in each of the TSAs, although the Virgin Islands declined to have its fourth-grade scores reported in 1992.

- How can NAEP ensure the fullest participation of states and schools? What burdens result from participation? What incentives might be offered?

- Are there cost savings or other efficiencies to be gained by building state NAEP as the foundation of national NAEP? Is this feasible?

- What would be the ideal NAEP assessment program, given no budget limitations? What choices are ahead, given funding realities?

The issues outlined above convey many of the complexities affecting NAEP. In addition there are calls for assessments in more subject areas, for annual assessments, for flexibility to allow states to choose the subjects and grade levels in which their students will be assessed, and for faster reports of results. The pace at which the government and contractors are working to respond to current and anticipated demands is already near the limit of existing resources.

It is to the credit of NAGB and NCES that NAEP has such appeal to its many constituents. However, it is increasingly difficult for NAEP to satisfy the sometimes competing demands placed on it and still retain its high level of technical integrity.

An Overview of This Report

This report is divided into three major sections: the first section (chapters 1 through 4) provides the background information and the specific findings of the Panel's evaluation of the 1992 TSA.

Chapter 2 examines the *administration* of the 1992 TSA. To begin, the Panel scrutinizes the sampling design for its adequacy and accuracy in representing the fourth and eighth graders in the participating states. The Panel considers whether the results were biased in any way by inaccurate sampling, low participation rates, or exclusion of important groups of students and asks whether the results of the 1992 trial are close to those that would have been obtained if every fourth- and eighth-grade student in participating states had been assessed. The Panel also examines how well the assessment was implemented: Were the same procedures followed in each assessment session? Were results comparable to those obtained for a matched sample of national NAEP subjects? How frequent were administration irregularities, and what was their impact on student performance? The chapter concludes with the state testing directors' perspectives on the administration and an analysis of the cost of the 1992 TSA to participating states.

The focus of chapter 3 is on the *content validity* of the assessment content. For mathematics, the framework for the 1992 assessment was the same one used for the 1990 assessment, and it was evaluated in the Panel's first report. Certain enhancements were, however, added to the mathematics item pool for 1992. The Panel therefore considers the adequacy of the mathematics item pool for the 1992 assessment, including the extent to which the items in the assessment reflect the NAEP Mathematics Framework and the more recent NCTM Standards. For reading, the Panel examines the consensus process that resulted in the 1992 Reading Framework. The

Panel considers whether the framework is comprehensive with regard to current conceptions of how reading *should* be taught and how it *is* taught. The Panel also considers the adequacy of the reading item pool and the fit between the items and the framework. Finally, the Panel reviews the process by which NAEP frameworks and items are developed and considers how it might be enhanced.

In chapter 4, the Panel examines the *impact of the TSA*, including the manner in which the TSA results were reported to the public, policymakers, and the press. Among the questions the Panel addresses in this chapter are: Were the TSA results useful to states? If so, in what ways? Do states feel that their participation in the TSA was worthwhile? How have the TSA results been used to advance educational reform in the states?

The second section (chapters 5 through 12) provides short, focused discussions of each of the important policy issues listed above. Chapter 5 introduces the section by presenting the set of *guiding principles* that the Panel has formulated to frame its analyses of policy issues.

In each of the remaining chapters, the Panel pays particular attention to a major policy issue and offers its *recommendations* related to that issue. Chapter 6 reviews the TSA and provides recommendations regarding the *reauthorization* of state NAEP. Chapter 7 reviews the current student *exclusions* from the assessment and appraises various possibilities for increasing inclusion, such as special versions of NAEP for children with disabilities and foreign-language versions of NAEP for children with limited proficiency in English. In chapter 8, the Panel's findings regarding achievement-level reporting are briefly reviewed, and alternatives for *reporting strategies* are offered. Chapter 9 reviews and re-evaluates the issues for and against lifting the *prohibition* against reporting results below the state level. Although the Panel has previously expressed its opposition to lifting the ban against reporting NAEP results below the state level, the issue continues to be proposed, so the Panel offers additional thoughts on this topic. The relationship between the *content* NAEP assesses and the evolving national content standards is the subject of chapter 10, and the feasibility of *linking* NAEP to state and international assessments is discussed in chapter 11. The issue of the burden for states participating in NAEP is addressed in chapter 12, with ideas for increasing the *participation* rates.

The last section of the report (chapter 13) deals with the *future* of NAEP. The Panel describes an ideal NAEP, the realities that impinge on it, and the principles and tradeoffs to be considered. Finally, an appendix provides *synopses* of the papers and studies that the Panel commissioned to obtain additional data for its evaluation.

2 *The Conduct of the 1992 Assessment*

For the 1992 TSA, the Panel evaluated several technical aspects of the assessment design and execution, replicating and expanding upon comparable analyses conducted by the Panel for the 1990 TSA.

- The Panel examined whether the sample chosen for study was capable of sustaining generalizations to how students in each of the participating states and territories are achieving. The sampling design and participation rates for both students and schools were evaluated, as were the weighting procedures used to adjust data for nonparticipation. For certain groups of students that were excluded in the sample design, the Panel also examined the appropriateness of the exclusions, the uniformity of the exclusion procedures across states, and the potential impact on state-by-state comparability of results.

- Several sources of data were analyzed to evaluate the uniformity and adequacy of the assessment administration. The performance of students in the TSA sample was compared to the performance of students in a comparable subsample of the 1992 national NAEP. Data collected by the contractor's quality control monitors and by Panel observers were analyzed to determine the frequency of administration irregularities and the relationship of these irregularities to student performance. Finally, student performance in classrooms visited by the quality control monitors was compared to performance in unmonitored classrooms.

- To address issues regarding the costs and benefits of the 1992 TSA, the Panel reviewed the results of a study of the costs to the states of participating in the TSA. And, as in 1990, state testing directors were asked about their experience implementing the TSA in their states and the perceived benefits of participation.

Table 2.1 compares the 1990 and 1992 TSAs in terms of the numbers of participating jurisdictions, schools, and students. As can be seen, the 1992 TSA included more grades and subjects and, in terms of numbers of students and numbers of sessions, was approximately three times as large as the 1990 TSA. For both TSAs, another major assessment—national NAEP—was conducted roughly in parallel by the same NAEP contractor (Westat), with most students tested during the same relatively brief time period.[1] Throughout this chapter, the Panel considers whether the increased size and scope of the 1992 assessment had an effect on its quality.

[1] For 1992, this period fell between February 3 and March 6, 1992, the inclusive dates for the TSA.

Table 2.1. ***Comparison of 1990 and 1992 Trial State Assessments***

	1990 TSA	1992 TSA
Grades assessed	8	4 and 8
Subject areas assessed	Mathematics	Mathematics and Reading
Number of participating states and territories	40	44
Approximate number of participating schools per state or territory	100	225
Total number of participating schools, after substitution	3,600	8,700
Total number of sessions	3,900	13,900
Total number of participating students	101,000	331,000
Percentage of sessions monitored	49.0	49.5–49.9
Number of assessment administrators trained	3,500	10,000

SOURCES: National Center for Education Statistics, *1990 Trial State Assessment: Summary of Participation Rates, Training, and Data Collection Activities* (Washington, D.C.: Author, September 1990) and National Center for Education Statistics, *1992 Trial State Assessment: Summary of Participation Rates, Training, and Data Collection Activities* (Washington, D.C.: Author, July 1992).

The Trial State Assessment Sample

The purpose of NAEP is to make objective information about student performance in selected learning areas available to policymakers at the national, regional, state, and local levels.[2] Furthermore, Public Law 100-297 requires that this information be "both representative and comparable."[3] To determine both the overall quality and the state-

[2] Public Law 100-297, Part C, Section 3402, (April 28, 1988), 344.

[3] Ibid.

to-state comparability of the TSA samples, the Panel evaluated the manner in which samples of schools and students were chosen for the assessment, the rates at which selected schools and students actually participated, and the weighting procedures used to compensate for those who did not participate.

The Panel also directed its attention to the major groups of students excluded in the design of the 1992 TSA: private school students and those public school students with either individualized education plans (IEP) or limited English proficiency (LEP) who were judged incapable of meaningful participation in the assessment. The Panel looked for external evidence that plausible numbers of IEP and LEP students had been excluded from the TSA overall, considered whether assessment administrators in different states had applied the criteria for excluding students in a consistent manner, and examined the potential impact of between-state differences in the proportions of private school, IEP, and LEP students on the comparability of state-by-state results.

Sampling and Participation of Schools

Sampling. The Panel commissioned a sampling expert to examine the methods by which schools were sampled for the 1992 TSA.[4] The sampling design called for approximately 100 schools per state for each of the three assessments: grade-4 reading, grade-4 mathematics, and grade-8 mathematics; significantly more schools were selected in states, such as Nebraska, with high proportions of small schools.[5] In most cases the two fourth-grade samples were overlapped so that, on average, only 128 fourth-grade schools were sampled per state: 13 assessing mathematics or reading only; 115 assessing both subjects. Approximately 2,500 students were assessed for each subject and grade combination in each state.

The sample of schools was taken from a list obtained from Quality Education Data (QED), Inc. Although the QED list is believed to be more current and to contain more collateral information than the NCES Common Core of Data (CCD), the QED list was checked against the CCD for each state. Procedures were also implemented to allow new schools that were on neither the QED nor CCD lists to be identified and sampled. The states' samples of schools were selected using a stratified sampling procedure—a systematic means for selecting schools within a state that ensures that the resulting sample will reflect the overall composition of the state on certain key variables. For the 1992 TSA, stratification was based on school size (defined as enrollment in the target grade), urbanicity,[6] percentages of black or Hispanic students,

[4] B.D. Spencer, "A Study of Eligibility Exclusions and Sampling: 1992 Trial State Assessment," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

[5] A small school has fewer than 20 students in the target grade. For further details on the sampling design and implementation, see E.G. Johnson, J. Mazzeo, and D.L. Kline, *Technical Report of the NAEP 1992 Trial State Assessment Program in Mathematics* (Washington, D.C.: National Center for Education Statistics, 1993).

[6] The urbanicity stratum included seven categories: large central city, mid-size central city, urban fringe of large city, urban fringe of mid-size city, large town, small town, or rural.

and estimated median family income. Within strata, schools were selected with probabilities proportional to their estimated fourth- or eighth-grade enrollment.

For states or other jurisdictions with small numbers of schools, and no or very few small schools, all schools were included in the sample with certainty. Thus, all the eligible fourth-grade schools in the District of Columbia, Delaware, Guam, and the Virgin Islands were taken into the fourth-grade sample with certainty. In the eighth grade, all the eligible schools were included for the District of Columbia, Delaware, Hawaii, Rhode Island, Guam, and the Virgin Islands.[7] While these jurisdictions represent the most extreme cases, in several slightly more populous states the same schools also must be sampled repeatedly for successive years of the TSA (including field tests), and large numbers of students must be sampled per school, particularly when more than one subject is assessed in a given grade.

The issue of school-sample overlap with other U.S. Department of Education studies is becoming increasingly relevant for states of all sizes. In 1992, in addition to the TSA at grades 4 and 8, NAEP collected national data from a number of distinct samples at grades 4, 8, and 12. And, at the same time, the U.S. Department of Education conducted a first-phase followup to a national longitudinal study of Chapter 1 children which involved substantial student assessment at grades 4 and 8. To avoid undue burden on individual schools, NAEP attempted to avoid overlap of school samples from different studies for the same age class. Distinct schools were also selected for the fourth- and eighth-grade state assessment samples within a state to the extent feasible.[8] Obviously, the smaller the state, the less likely that school-sample overlap could be completely avoided.

NAEP is currently exploring options for reducing the numbers of small schools sampled for the TSA and, in some cases, the numbers of students sampled per school. This issue, which is favored by state testing directors in the most heavily impacted states, is discussed in more detail in chapter 12.

The TSA school-sample design met the standards for sample designs used in high quality, major national education surveys. It was executed with care and due attention to detail. This favorable result replicates the results of the review of the sample design conducted in 1990. However, as noted above, the design placed greater burdens on states with proportionately fewer schools or fewer students.

Participation rates. Recruitment of schools for the TSA was carried out by state coordinators, working from lists provided by the NAEP contractor. When schools declined to participate, substitute schools were identified by the contractor and recruited by the state coordinator. Substitute schools were chosen to be as similar as possible to the refusing school with regard to urbanization, percent black enrollment, percent Hispanic enrollment, total grade-level enrollment, and median income. In cases where no suitable substitute could be found among those schools not sampled (most often because all or most schools were included in the original sample), a school already in the sample was asked to conduct a double session, with one session serving as a substitute for students in the refusing school.[9] Average participation rates

[7] Johnson et al., op. cit., 69.

[8] Ibid., 67.

[9] Ibid., 77-78.

for originally sampled schools for both the 1990 and 1992 TSAs are shown in table 2.2. Participation rates for national NAEP are included for comparison. National NAEP rates that are directly comparable to the TSA rates (because they are based only on public-school cooperation rates) are shaded. Between the 1990 and 1992 grade-8 state assessments, school participation rates dropped from 94 percent to 89 percent. Corresponding national NAEP rates for grade 8 were 87 percent in 1990 and 88 percent in 1992. The higher TSA participation rate in 1990 may reflect the "novelty effect" of the first state assessment.

Table 2.2. ***Initial school participation rates (percent of originally sampled schools participating) in recent national and state NAEP assessments***

	Grades 3-4	Grades 7-8	Grades 11-12
1983-84 national NAEP	89	90	84
1985-86 national NAEP winter sample	87	84	n/a
1985-86 national NAEP spring sample	89	89	81
1990 national NAEP	88	87	81
1990 TSA	n/a	94	n/a
1992 national NAEP	86	88	81
1992 TSA	89	89	n/a

SOURCES: A.E. Beaton et al., *Implementing the New Design: The NAEP 1983-84 Technical Report* (Washington, D.C.: National Center for Education Statistics [NCES], 1987), 87; A.E. Beaton et al., *Expanding the New Design: The NAEP 1985-86 Technical Report* (Washington, D.C.: NCES, 1988), 71; I.V.S. Mullis, J.A. Dossey, E.H. Owen, and G.W. Phillips, *The STATE of Mathematics Achievement* (Washington, D.C.: NCES, June 1991), 453; S.L. Koffler et al., *The Technical Report of NAEP's 1990 Trial State Assessment Program* (Washington, D.C.: NCES, April 1991), 55; and I.V.S. Mullis, J.A. Dossey, E.H. Owen, and G.W. Phillips, *NAEP 1992 Mathematics Report Card for the Nation and the States* (Washington, D.C.: NCES, April 1993), 316-318, 345.

Although *average* school participation rates for the 1992 TSA were 89 percent before substitution, these initial participation rates were less than 85 percent in a third of the participating states, as shown in table 2.3. In addition, two states (Illinois and Washington) withdrew from the 1992 TSA during school recruitment due to high numbers of refusing schools. This is a major change from 1990, when only *two* states had initial school participation rates below 85 percent.

After the substitute schools were included in the sample, final participation rates for schools were much higher.[10] However, even when substitutes are matched as carefully as possible to the refusing schools, it is not possible to be certain that all potential sources of bias have been eliminated. Initial, before-substitution participation rates are a much more reliable indicator of absence of bias in the resultant sample and are therefore given much greater weight in the Panel's evaluation.

Table 2.3. ***States with before-substitution school participation rates below 85 percent: 1990 and 1992 TSAs (weighted percentage participating in parentheses)***

1990 TSA grade-8 mathematics	1992 TSA grade-8 mathematics	1992 TSA grade-4 mathematics	1992 TSA grade-4 reading
Illinois (78)	Alabama (66)	Alabama (75)	Alabama (76)
Oklahoma (78)	Indiana (79)	Idaho (84)	Idaho (82)
	Maine (62)	Indiana (76)	Indiana (77)
	Massachusetts (83)	Maine (57)	Maine (58)
	Michigan (78)	Michigan (83)	Michigan (83)
	Minnesota (81)	Minnesota (82)	Minnesota (81)
	Nebraska (75)	Nebraska (80)	Nebraska (76)
	New Hampshire (80)	New Hampshire (69)	New Hampshire (68)
	New Jersey (69)	New Jersey (76)	New Jersey (76)
	New Mexico (77)	New Mexico (75)	New Mexico (76)
	New York (81)	New York (78)	New York (78)
	North Dakota (78)	North Dakota (73)	North Dakota (70)
	Ohio (77)	Ohio (79)	Ohio (78)
	Oklahoma (82)	Pennsylvania (84)	Rhode Island (83)
	Pennsylvania (81)	Rhode Island (83)	

SOURCES: I.V.S. Mullis, J.A. Dossey, E.H. Owen, and G.W. Phillips, *The STATE of Mathematics Achievement* (Washington, D.C.: National Center for Education Statistics [NCES], June 1991), 437; I.V.S. Mullis, J.A. Dossey, E.H. Owen, and G.W. Phillips, *NAEP 1992 Mathematics Report Card for the Nation and the States* (Washington, D.C.: NCES, April 1993), 316-318.

[10] At grade 4, final rates were all above 90 percent except in Maine (71 percent), Nebraska (87), New Hampshire (80), and New Jersey (82); at grade 8, the exceptions below 90 percent were Maine (84 percent), Nebraska (85), New Jersey (78), and New York (83).

In 1990, Spencer's analysis of the effects of school nonparticipation within a single state (Illinois) did not yield any major differences in terms of mathematics achievement.[11] However, more recent Panel analyses using participating states as the unit of analysis found a consistent relationship between before-substitution participation rates and mean state scores on NAEP (see table 2.4). Lower before-substitution participation rates were slightly associated with higher mathematics achievement in 1990, and this association was stronger in 1992. The fact that the finding was replicated in both the fourth- and eighth-grade samples suggests that the result is not due to chance. Furthermore, it does not appear that the finding is due to one or two aberrant states.

Because the analyses that led to these findings were performed only shortly before this report went to press, the Panel has not had the opportunity to undertake additional analyses that might help to explain the association between initial participation rates and states' achievement scores. The Panel intends to give high priority to further investigation of this issue during 1994, however.

Table 2.4. ***Correlations between weighted, before-substitution school participation rates and average state proficiency***

1992 TSA			1990 TSA
Grade-4 reading	Grade-4 mathematics	Grade-8 mathematics	Grade-8 mathematics
-.447 (n = 43)	-.367 (n = 43)	-.361 (n = 44)	-.286 (n = 40)

In light of the large number of states with initial participation rates below 85 percent, as well as the finding that lower initial participation rates were reliably associated with higher state scores, the Panel concludes that initial participation rates for some states were too low for accurate reporting of their 1992 TSA results. The Panel gives further consideration to the issue of adequate initial participation rates in chapter 6.

Sampling and Participation of Students

Sampling. In each selected school, a sample of about 30 students was drawn from a list that schools provided by November 15, 1991, 2 1/2 months before the assessments

[11] Spencer performed an analysis to test for the effect of school nonparticipation in the 1990 TSA in Illinois. Mean scores from the 1990 Illinois Goals Assessment in eighth-grade mathematics were obtained for all schools eligible for the 1990 TSA in Illinois and were transformed to standardized scores such that the state mean score was zero. The average score for the 105 originally selected TSA schools was -0.077. After deletion of the 23 nonparticipating schools and addition of the 19 replacement schools, the recomputed average was -0.050.

began in February 1992.[12] Between the time of sample selection and the time of assessment, some students who were originally selected moved to other schools. To allow for appropriate sampling of students changing schools during this time period, each school was asked to provide a supplementary list including the names of all students who had moved *into* the school since the previous list was produced. Students were then sampled from the supplementary list at the same rate as the original sample, and those selected were added to the sample.

Analyses of the 1990 TSA data showed that substantially fewer students were added from these supplemental lists than were lost from the original lists.[13] This finding was replicated in 1992. Table 2.5 shows that although about 5 percent of students withdrew from the TSA sample, only about 3 percent were added by the supplementary sampling procedure.[14] Thus, only 60 percent as many students were captured by supplementary sampling as were lost by withdrawals, presumably as a consequence of dropouts, delays in re-enrollment, seasonal migration to the families' countries of origin, and other temporary or permanent withdrawals from public school.

Table 2.5. ***Percent of withdrawals in the 1990 and 1992 Trial State Assessments compared to percentage of students added by the supplementary sampling procedure***

	Percentage of students withdrawn	Percentage of students added by supplementary sampling	Difference between percentage withdrawn and percentage added
1990 TSA grade-8 mathematics	4.9	2.9	2.0
1992 TSA grade-4 mathematics	4.5	3.1	1.4
1992 TSA grade-8 mathematics	4.8	2.8	2.0

SOURCES: National Center for Education Statistics, *1990 Trial State Assessment: Summary of Participation Rates, Training, and Data Collection Activities* (Washington, D.C.: Author, September 1990) and National Center for Education Statistics, *1992 Trial State Assessment: Summary of Participation Rates, Training, and Data Collection Activities* (Washington, D.C.: Author, July 1992).

[12] In the smallest jurisdictions, however, the number of students was much higher. For example, in Delaware and the District of Columbia, up to 120 students per school were selected.

[13] The National Academy of Education, *Assessing Student Achievement in the States* (Stanford, CA: Author, 1992), 9.

[14] National Center for Education Statistics, *Summary for All States, Training and Data Collection Activities, 1992 Trial State Assessment* (Washington, D.C.: Author, May 1992, draft).

To a considerable extent, high student mobility is a function of high family mobility and thus associated with unemployment or underemployment, shortage of low-cost housing, immigrant status, or family problems.[15] Such conditions are known to correlate with lower academic performance. Furthermore, analyses of the 1992 TSA data showed that the students who, by self report, had changed schools at least once within the past 2 years scored considerably lower than did students who had not changed schools. *Because of the bias introduced by withdrawn students who are not represented in the supplemental sample, it is likely that the TSA data overestimate overall performance in states with high proportions of mobile students and dropouts.* Further investigation is necessary in order to better understand the characteristics of withdrawn students and how this class of students can best be represented in the NAEP sample.

Participation rates. TSA procedures require that a makeup session be given whenever three or more students are absent from the initial assessment session; however, relatively few of the absentee students attended the makeup sessions. Nevertheless, very good student participation rates were obtained for all states except Oklahoma.[16] Grade-4 participation rates for selected students who were not excluded for cause (e.g., student not in target grade, IEP or LEP) ranged from 93 percent to 97 percent except in Oklahoma, where the rate was between 84 and 85 percent. Rates for grade 8 ranged from 93 percent to 97 percent except in Oklahoma, where the rate was 84 percent. These 1992 rates compare favorably with the 1990 TSA, in which student participation rates ranged from 92 percent to 96 percent, except in the District of Columbia (88 percent) and Oklahoma (80 percent).

Weighting

In a survey as complex as the TSA, sampling weights are required in order to make valid inferences from the student samples to the respective populations from which they were drawn. In 1992, these weights contained three components. First, a base weight was established that was the inverse of the overall probability of selection of the sampled student. The base weight incorporated the probability of selecting a school and the student within a school, and accounted for the impact of procedures used to keep to a minimum the overlap of the state school sample with the NAEP national sample and the sample for the National Longitudinal Study of Chapter 1 Children. The base weight was then adjusted for two sources of nonparticipation—school level and student level. These latter adjustments increased (1) the weights of students from schools similar to schools that did not participate and (2) the weights of students similar to students within participating schools who did not attend the assessment session (or a makeup session) as scheduled.[17]

[15] L.S. Miller, M. Fredisdorf, and D.C. Humphrey, *Student Mobility and School Reform* (New York, NY: Council for Aid to Education, 1992), 1.

[16] Oklahoma was the only state that required parental permission forms on a statewide basis.

[17] Johnson et al., op. cit., 129.

Adjustments for nonparticipation were performed separately for each state. For nonparticipating schools, weights were reassigned to schools that were matched on the previously mentioned sampling strata of urbanization, percent minority, and estimated median family income. Weights for nonparticipating students within participating schools were reassigned on the basis of these same school-level characteristics, plus the monitor status of the student's assigned session (monitored or unmonitored) and the age of the student relative to others in his or her grade level (older or not older). Thus, only one student-level characteristic—a dichotomous age variable—contributed to the reassignment of weights for nonparticipating students. This appears to be unacceptably crude given that it is well known that participating and nonparticipating students tend to differ on many characteristics, including overall test performance. Not only is there evidence from other educational surveys (such as High School and Beyond) that nonparticipating students tend to be lower scorers on tests (when they eventually do take the tests), but also in the 1992 TSA, absentee students who attended makeup sessions generally did not perform as well as students who attended their originally scheduled sessions.[18] Furthermore, participation rates are known to vary by gender and ethnicity, characteristics that correlate with mean test performance.

The Panel's evaluation of the 1992 TSA found that the weighting procedures used to compensate for different selection rates among different categories of students was reasonably straightforward and that the adjustments for school nonparticipation were acceptable. The adjustments for student nonparticipation, however, were based on criteria that were unlikely to be very effective in removing student-level nonresponse bias. A more precise correction would require additional information on the demographic characteristics of nonparticipating students.

Excluded Students

Three major groups of students were excluded from the design of the 1992 TSA: students enrolled in private schools and IEP or LEP students in public schools, if they were judged incapable of participating meaningfully in the assessment. Nationwide, about *one in six* students belongs to one or more of these groups (private school, LEP, IEP), and states vary widely in the proportions of private school, IEP, and LEP students in their populations. These variations have potentially important implications for state-to-state and state-to-national comparisons. Furthermore, the NAEP design calls for the exclusion criteria for IEP and LEP students to be applied locally, by personnel in the participating schools. If these exclusion criteria are not applied consistently across states, the resulting samples of students may not adequately represent each state's testable population. For all of these reasons, the sampling expert commissioned by the Panel examined the effects of exclusion in some detail.

[18] Compared to nonmakeup students in grade-8 mathematics, there were about 18 percent fewer makeup students at or above the basic achievement level, about 9 percent fewer at or above the proficient level, and about 1 percent fewer at or above the advanced level. The proportions of nonmakeup students scoring at or above these levels were 61 percent, 21 percent, and 2 percent, respectively.

Private school students. Private school students typically make up the largest excluded group in each state. Nationally, it is estimated that about 11 percent of fourth-grade students and 10 percent of eighth-grade students are enrolled in private schools, but these percentages vary widely across states. For example, at grade 4, only about 1 percent of students in Utah are enrolled in private schools, compared with 20 percent of students in the District of Columbia. In grade 8, the percentages of students enrolled in private schools range from 2 percent (Wyoming) to 25 percent (District of Columbia). Table 2.6 shows differences in performance between public and private school students at grades 4 and 8 on the 1992 national NAEP in mathematics. At both grades, private school students tended to perform better than public school students. For a given state, the magnitude of the effect of excluding private school students depends on the number of students enrolled in private schools and on the size of the performance differences between public and private school students.

Table 2.6. ***Performance of private and public schools on the 1992 national NAEP in mathematics***

	Average proficiency by type of school		
	Public	Catholic	Other private
Grade 4	217	227	226
Grade 8	266	277	284

SOURCE: I.V.S. Mullis, J.A. Dossey, E.H. Owen, and G.W. Phillips, *NAEP 1992 Mathematics Report Card for the Nation and the States* (Washington, D.C.: National Center for Education Statistics, April 1993), 149.

IEP and LEP students. The Panel's study of the exclusion of IEP and LEP students examined two components of the exclusion process: (1) the rates at which students were *classified* or identified as having an individualized education plan or limited English proficiency, and (2) the rates at which these classified students were subsequently *excluded* from the assessment.

For each session, assessment administrators were asked to record each student's IEP or LEP status, whether or not the student was to be excluded from the assessment. They were then asked to determine and record IEP and LEP exclusions based on the following instructions: "Identify students who are incapable of taking the assessment according to the criteria provided for excluding students [shown in figure 2.1]. Remember that the original list from which the sample was selected was to include *all* students in the fourth grade regardless of their ability to be assessed. Now is the time to identify those IEP or LEP students who are incapable of taking the assessment according to NAEP criteria."[19]

[19] Westat/NAEP, *Manual for Assessment Administrators, Fourth Grade, 1992 Trial State Assessment* (Rockville, MD: Author, 1992).

Figure 2.1. ***Criteria for excluding students from NAEP***

The intent is to assess all selected students. Therefore, all selected students who are capable of participating in the assessment should be assessed.

Some of the students identified on the Administration Schedule as Limited English Proficient (LEP) or as having an Individualized Education Plan (IEP) may be incapable of participating meaningfully in the assessment. The Assessment Administrator, with the advice of staff members knowledgeable about the IEP/LEP students, may exclude such students, as described below.

1. A student identified on the Administration Schedule as LEP may be excluded from the assessment if he/she:

   - Is a native speaker of a language other than English;

   AND

   - Has been enrolled in an English-speaking school (not including a bilingual education program) for less than two years;

   AND

   - Is judged to be incapable of taking part in the assessment.

2. A student identified on the Administration Schedule as having an IEP or equivalent classification may be excluded from the assessment if:

   - The student is mainstreamed less than 50 percent of the time in academic subjects and is judged incapable of participating meaningfully in the assessment.

   OR

   - The IEP team or equivalent group has determined that the student is incapable of participating meaningfully in the assessment.

3. IEP/LEP students meeting the above criteria should be assessed if, in the judgment of school staff, they are capable of taking the assessment.

   WHEN THERE IS DOUBT, INCLUDE THE STUDENT.

SOURCE: Westat/NAEP, *Manual for Assessment Administrators, Fourth Grade, 1992 Trial State Assessment* (Rockville, MD: Author, 1992).

Overall, about 9 percent of the sampled students in each grade and subject were classified as IEP, and 5 percent were excluded based on their IEP status. For LEP, the corresponding figures were 6 percent classified and 3 percent excluded at grade 4, 4 percent classified and 2 percent excluded at grade 8. Percentages across states varied: IEP exclusions ranged from about 2 percent in North Dakota and the Virgin Islands to 7 or 8 percent in several states and the District of Columbia. LEP exclusions ranged from 0 to 2 percent in most states, but they reached 5 percent for grade 8 and 10 to 11 percent for grade 4 in California. The state-by-state variations in *LEP* classification and exclusion rates were generally consistent with expectations based on state demographics. No similarly self-evident reason for the state-by-state variations in *IEP* rates was apparent, however.

In general, state exclusion rates were consistent across the three 1992 TSA assessments. Rates were also similar for the 1990 and 1992 grade-8 mathematics TSAs. Figures 2.2 through 2.4 illustrate the range of state-by-state variation in exclusion and inclusion rates for IEP and LEP students combined. These figures show the percentage of the originally sampled students for each state who were classified as either IEP or LEP, then subsequently either included or excluded.

Figure 2.2. ***State-by-state rates of inclusion and exclusion of IEP and LEP students, 1992 TSA, grade-8 mathematics***

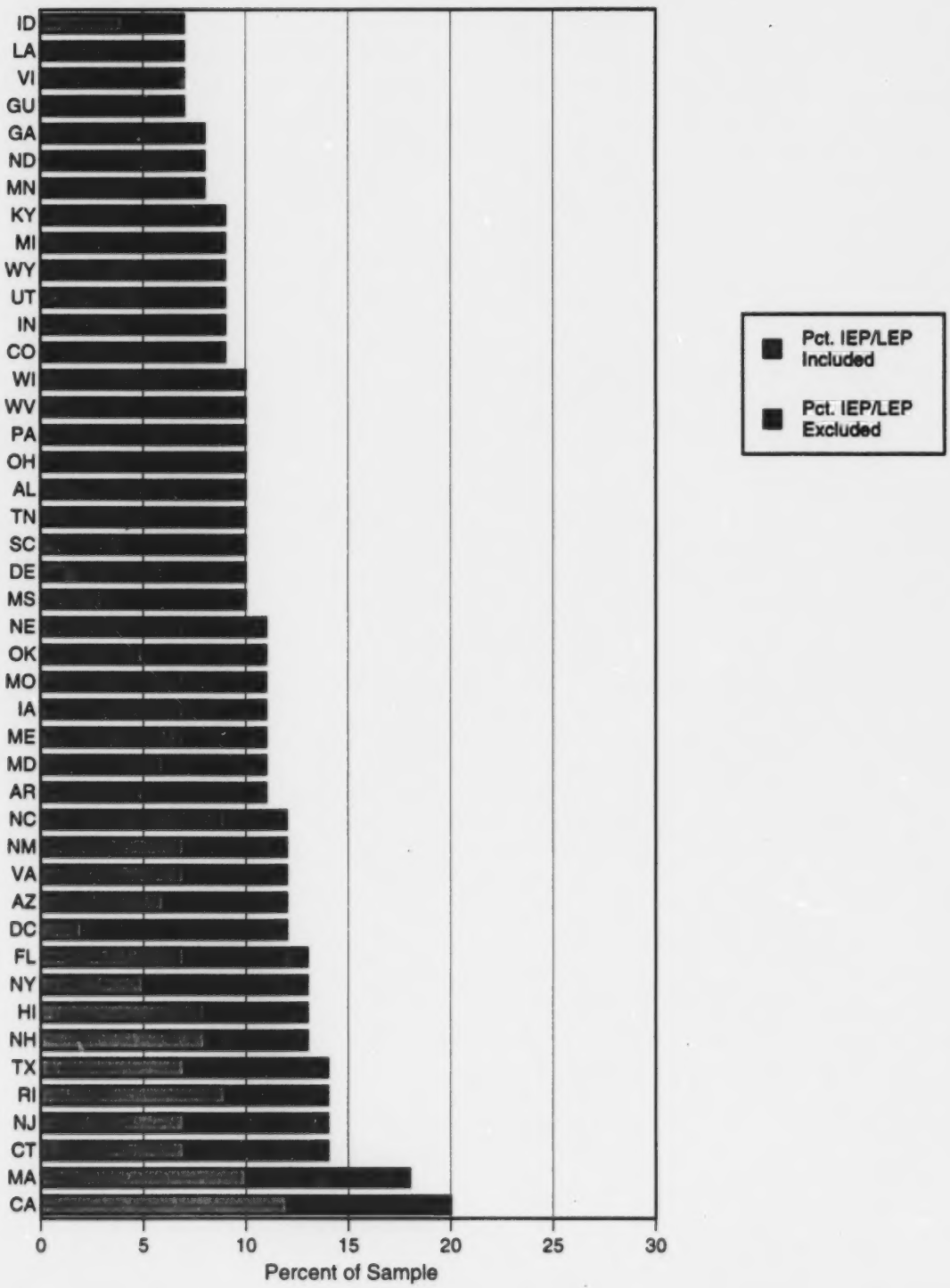


Figure 2.3. ***State-by-state rates of inclusion and exclusion of IEP and LEP students, 1992 TSA, grade-4 mathematics***

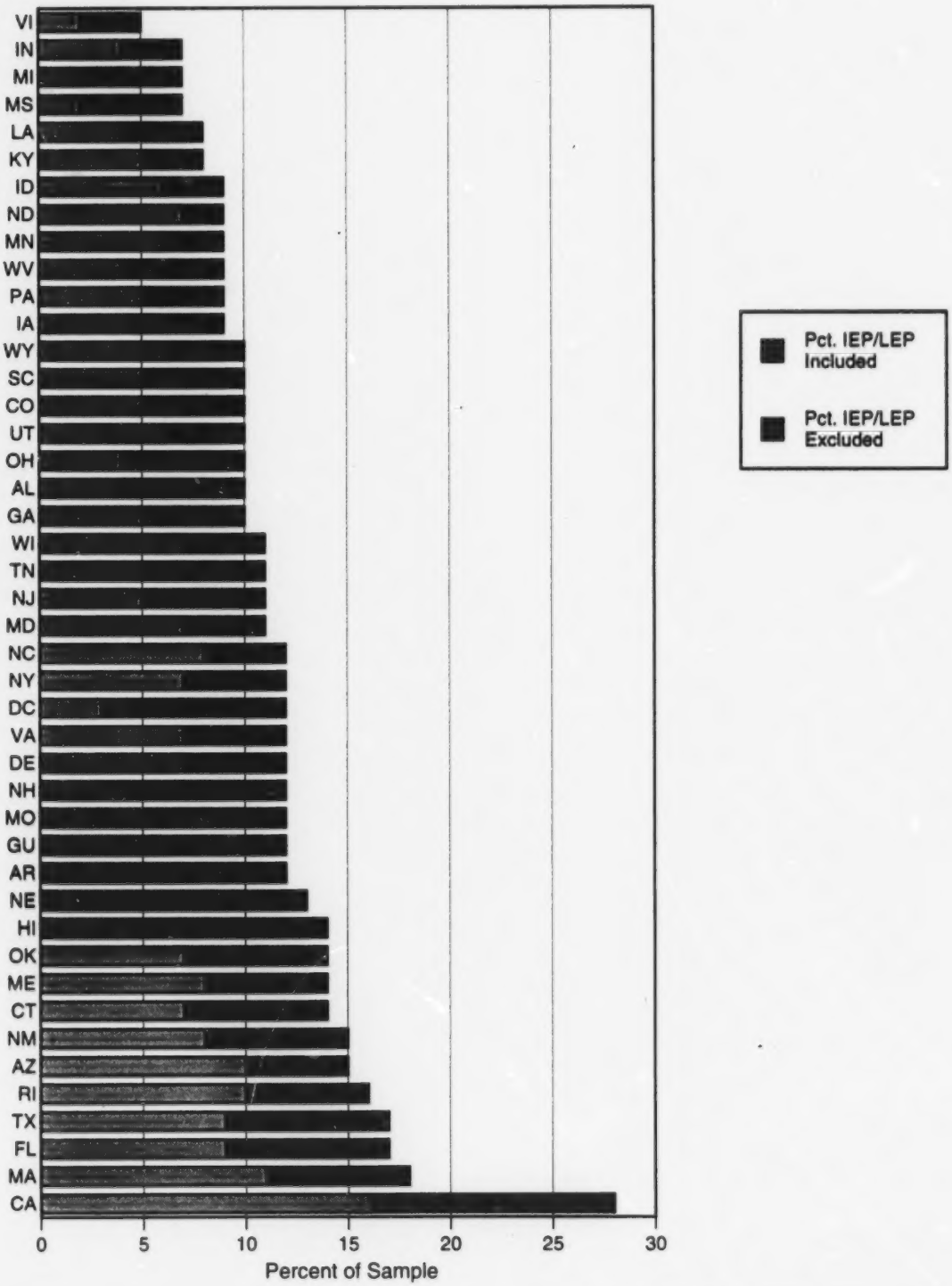
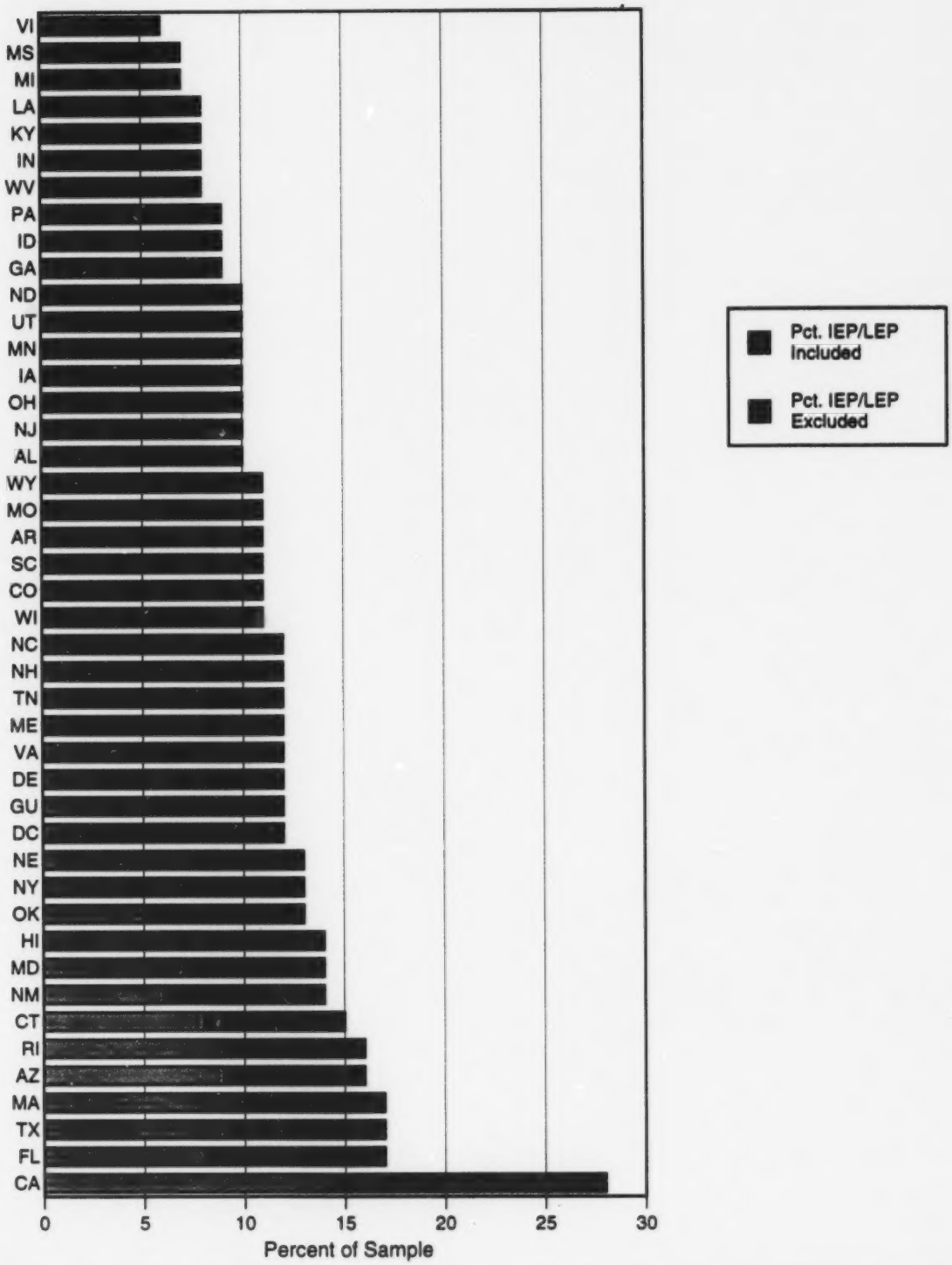


Figure 2.4. ***State-by-state rates of inclusion and exclusion of IEP and LEP students, 1992 TSA, grade-4 reading***



In order to assess the reasonableness of the IEP and LEP classification and exclusion rates observed in the 1992 TSA, these rates were compared with various external sources. Table 2.7 compares classification and exclusion rates for the 1990 and 1992 TSAs against national NAEP. Overall, the rates seem reasonably consistent across the two surveys, although the classification rates are somewhat higher in the TSA and the exclusion rates (for 1992) are slightly lower. The higher classification rates in the TSA (and the greater consistency of the classification rates between 1990 and 1992) may reflect the fact that the local school personnel who serve as assessment administrators for this survey are more familiar with the sampled students, or have better access to school records, than do the contractor staff who administer national NAEP.

Table 2.7. ***Comparison of IEP and LEP classification and exclusion rates for 1990 and 1992 Trial State Assessments and national assessments***

	IEP and LEP		IEP		LEP	
	% classified	% excluded	% classified	% excluded	% classified	% excluded
1992 grade-4 reading TSA	15	7	9	5	6	3
1992 grade-4 reading national assessment	12	8	9	6	4	3
1992 grade-4 mathematics TSA	14	6	9	4	6	3
1992 grade-4 mathematics national assessment	12	8	9	6	4	3
1992 grade-8 mathematics TSA	13	6	9	5	4	2
1992 grade-8 mathematics national assessment	12	7	9	5	3	2
1990 grade-8 mathematics TSA	13	7	9	5	3	2
1990 grade-8 mathematics national assessment	8	6	6	5	2	1

SOURCE: B.D. Spencer, "A Study of Eligibility Exclusions and Sampling: 1992 Trial State Assessment," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

TSA IEP and LEP classification rates were also compared with available national statistics. For IEPs, the comparison was to state-by-state figures provided by the Department of Education based on participation in federal programs under the Individuals with Disabilities Education Act (IDEA). In general, the TSA classification rates tended to be similar to the IDEA rates but slightly lower. Unexplained variations in state-by-state IEP rates were observed in the IDEA data as well as in the TSA, and the correlation between individual states' TSA and IDEA classification rates was only low to moderate. Consequently, the analyses did not shed a great deal of light on the reasonableness of the observed state-by-state differences in TSA IEP classification rates.

LEP classification rates were compared with figures derived from the Survey of States' Limited English Proficient Persons and Available Educational Services, School Year 1991-1992. There was general agreement between the two sources except in a few states. Large discrepancies were observed in New Mexico, North Dakota, and Utah, but appeared to be due to the classification of Native Americans as LEP in the Educational Services survey, regardless of their proficiency in English.

Panel staff also analyzed assessment data to search for indications that schools applied the NAEP criteria for exclusion consistently. It appears that some schools may have excluded all IEP and LEP students from the TSA outright, because in a small percentage of sessions (4 percent at grade 4, and 7 percent at grade 8) large numbers of students (five or more) who were classified as IEP or LEP on the administration schedule were categorically excluded from the assessment.[20] As discussed in chapter 7, the Panel will be undertaking a more extensive investigation of the appropriateness of local exclusion decisions in conjunction with the administration of the 1994 TSA.

Finally, the Panel examined distributions of scores for the assessed (included) IEP and LEP students on the 1992 mathematics assessments for both grades 4 and 8. There was a large amount of overlap between the performance scores for IEP, LEP, and non-IEP or LEP students, and student-by-student response patterns on constructed-response questions indicated that the IEP and LEP students performed at about the same level as a low-scoring comparison group. These findings suggest that, as a group, the included IEP and LEP students were, in fact, able to participate meaningfully in the assessment.

Effects of exclusions. In 1990, statistical (sensitivity) analyses were conducted to determine the extent to which the exclusion of private school students might have affected state performance on the TSA, under varying assumptions about the size of the differences in public school versus private school student performance.[21] The conclusion from these analyses was that results for some participating states would have changed noticeably (by as much as 2.5 state ranks) if private schools had been included in the TSA. Based on this finding, the Panel recommended that state NAEP test private as well as public school students, a recommendation that is being implemented for the 1994 TSA. For 1992, the Panel's sensitivity analyses were repeated for the TSA in mathematics, but using assumptions more closely tied to

[20] E. Hartka, "A Study of Students Excluded from the 1992 National Assessment of Educational Progress Trial State Assessment," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

[21] B.D. Spencer, "Eligibility/Exclusion Issues in the 1990 Trial State Assessment," in *Assessing Student Achievement in the States: Background Studies* (Stanford, CA: The National Academy of Education, 1992).

actual student performance—namely that differences between public and private schools would be constant for all states within a NAEP region (Northeast, Southeast, Central, and West) and equal to the mean public/private school difference observed in that region on the 1992 national NAEP. The calculations may have underestimated the changes in rankings that would actually have occurred if private school students had been tested, because they ignore interstate variations in public/private school differences within a region. Nevertheless, the results of the Panel's analyses suggested that including private school students would have affected state means enough to change state ranks by a modest amount.[22]

The 1992 sensitivity analyses were also expanded to test for effects of including IEP and LEP students under two types of assumptions: (1) that excluded IEP or LEP students would have scored, on the average, at the level of the tested IEP or LEP students (the "optimistic" assumption), or (2) that IEP or LEP students would have performed at the very bottom of the distribution of observed scores (the "pessimistic" assumption). Results are shown in table 2.8. When all IEP and LEP and private school students are included, the changes in rank for some states are substantial. For example, in California (which has a high percentage of LEP students) the state rank for eighth-grade mathematics (29 out of 44) would decline to 31 (-2) under the optimistic assumption and to 37 (-8) under the pessimistic assumption. Conversely, states with higher proportions of private school students, such as Delaware, could expect their state's rank to improve. The analyses suggest that for fourth-grade mathematics, Delaware's rank would have improved from 24 out of 44 to 20 (optimistic IEP or LEP) or 19 (pessimistic IEP or LEP).

[22] The average change in rank was less than 1. The maximum change in rank at grade 8 was 3 (for Oklahoma); at grade 4 it was 4 (for Delaware).

Table 2.8. ***Illustrative effects of including excluded students in the 1992 TSA***

		Optimistic assumptions		Pessimistic assumptions	
Included group	Subject and grade	Average difference in ranks	Maximum difference in ranks	Average difference in ranks	Maximum difference in ranks
IEP	Mathematics grade 8	0.33	+2 for NC -2 for WV	1.90	+6 for NC
IEP	Mathematics grade 4	0.43	-2 for CO, WI +2 for MD	1.76	-5 for OK
LEP	Mathematics grade 8	0.24	-2 for FL	1.57	-9 for CA
LEP	Mathematics grade 4	0.48	-3 for RI	1.71	-6 for RI
IEP/LEP	Mathematics grade 8	0.43	+3 for NC	2.10	+9 for NC
IEP/LEP	Mathematics grade 4	0.52	-3 for RI	2.19	-6 for TX
IEP/LEP and private	Mathematics grade 8	1.00	-6 for OK	2.29	+9 for DE
IEP/LEP and private	Mathematics grade 4	1.00	+4 for DE	2.19	-8 for TX

NOTE: + indicates an improvement of state rank under this assumption - indicates a decline in relative position for a state

Summary. *The Panel's evaluation of the excluded-student data found that rates of IEP and LEP exclusion were generally consistent between the 1990 and 1992 TSAs. Furthermore, state-by-state estimates of proportions of IEP and LEP students were in general agreement with estimates of these populations obtained from independent sources. Further study is needed to determine whether the IEP and LEP exclusion criteria were applied consistently and appropriately by local assessment administrators.*

The two types of exclusion examined here (private school, IEP, and LEP students) have opposite effects on a state's performance. For the 1992 mathematics assessment, data from national NAEP suggest that private school students could be expected to perform

better than public school students. On the other hand, IEP and LEP students who were included in the TSA in mathematics tended to score near the bottom of the distribution, although there was considerable overlap with the performance of non-IEP and LEP students. Statistical sensitivity analyses carried out by the Panel suggested that the inclusion of all private school, IEP, and LEP students in the 1992 mathematics assessment would have markedly changed the mean performance and rank for a number of states.

The Administration of the 1992 TSA

The design of the trial state assessments requires that responsibility for the conduct of the assessments be shared by the NAEP contractor and by state and local personnel in the participating states. For example, school and student samples are drawn by the contractor, but state coordinators in each state bear primary responsibility for school recruitment and scheduling. The actual administration of the assessment, along with certain related activities, falls under the purview of the local assessment administrators who are, in most cases, faculty or staff members recruited from the sampled schools. Contractor staff provide training for the assessment administrators, contact the administrators shortly before the scheduled date of the assessment to confirm receipt of materials and other assessment arrangements, and, for 50 percent of the sessions, travel to the assessment sites on the day of the assessment to monitor the session administration. These arrangements differ from national NAEP, in which contractor staff have primary responsibility for all aspects of the assessment.

For the 1990 TSA, the Panel concluded that the assessment was carefully planned and implemented and that the training was adequate for the assessment. No statistically significant differences in student performance were observed between subjects in monitored sessions and sessions that were not monitored. However, the Panel found evidence that several types of administration irregularities, when they occurred, were significantly associated with lower average performance scores on the TSA. In addition, TSA subjects performed slightly but reliably better than a matched sample of national NAEP subjects. Several recommendations for future refinement of assessment procedures were made in the Panel's report, including suggestions that monitoring of sessions be continued at the current rate (50 percent of sessions) and that future training of assessment administrators continue to emphasize the need for uniform administration procedures.[23]

In 1992, the Panel once again looked at evidence for the uniformity of assessment administration across sessions and states. The mean and variance of student scores in mathematics and reading were compared between the TSA and a matched sample drawn from national NAEP as well as between monitored and unmonitored TSA sessions. In addition, student scores were examined in relation to the contractor's (quality control monitors') observations of assessment sessions. These observations were used to determine whether irregularities of administration were associated with increases or decreases in performance. Finally, Panel staff observed a small number of training sessions and assessment sessions to gather firsthand information about potential problems and areas needing improvement.

[23] D.H. McLaughlin, F.B. Stancavage, J.G. Chambers, E. Hartka, and K. Ercikan, "Evaluation of the Implementation of the 1990 Trial State Mathematics Assessment," in *Assessing Student Achievement in the States: Background Studies* (Stanford, CA: The National Academy of Education, 1992).

Trial State Assessment and National NAEP Comparisons

The great value of state NAEP is that it provides states with uniform student data that are directly comparable to national data; consistency between the state and national programs, therefore, is of utmost importance. A strong test of this consistency, and, by extension, of the adequacy of the state assessment administration conditions, is to compare overall performance for each grade and subject of the TSA against the overall performance of a matched sample drawn from national NAEP.[24] Of course, if performance were not comparable, this would not necessarily be due to differences in the administration. Students could be more or less motivated to perform in the TSA, for example, irrespective of how the administration was handled. It is less likely, however, that there could be substantial differences in the conditions of administration that would not be reflected in differences in overall performance.

In 1990, the state-to-national comparison revealed that, for the one grade and subject tested, TSA subjects performed slightly but reliably better than their counterparts in the national sample. Fortunately, this finding was not replicated in the current evaluation. *No statistically significant differences in overall performance were observed between the TSA and matched national samples for any of the subjects or grades assessed in the 1992 TSA.*

Training and Monitoring of Assessment Administrators

In 1992, Westat, the NAEP contractor for administration, trained nearly 10,000 local assessment administrators using a video presentation accompanied by a scripted trainer's guide and practice exercises. The video presentation was designed to facilitate uniform training procedures across all states. The entire training session generally ran for about 3 1/2 hours, substantially shorter than the 5- to 6-hour training sessions presented in 1990. Session length was reduced in response to feedback received from state and local personnel after the 1990 TSA.[25]

About 400 quality control monitors were first trained in 2-day sessions and then assisted with the training sessions for local assessment administrators. In this manner, the quality control monitors became thoroughly familiar with the assessment administrator's responsibilities. Panel staff attended one of the training sessions for quality control monitors and also observed a training session for assessment administrators. These observers noted that the training was of very high quality and was conducted in a professional manner by Westat.

[24] The drawing of a matched sample is necessary because national NAEP is designed to produce estimates for the entire nation and not just for the aggregation of states that participated in the TSA.

[25] Johnson et al., op. cit., 95.

Irregularities in monitored and unmonitored sessions. Despite the reduced training time for assessment administrators, irregularities reported by the quality control monitors (who completed standardized protocols for each session they observed) were both infrequent overall and somewhat less frequent than in 1990 (table 2.9). Furthermore, while there is no direct record of the irregularities that occurred in *unmonitored* sessions, it is very reassuring that student performance (both mean levels of performance and variability of performance within sessions, between schools, and between states) was essentially the same in monitored and unmonitored classrooms. Just as with the state-to-national comparisons discussed above, comparability of overall performance between monitored and unmonitored sessions provides a very good indicator that substantial differences in administration did not occur. (In these analyses, variability of performance was scrutinized even more closely than mean performance because irregularities affecting performance would likely occur in only a subset of unmonitored classrooms.)

Table 2.9. ***Percentage of monitored sessions in which each of the four most common administration irregularities was observed***

Most common irregularities	1992 TSA			1990 TSA
	Grade-4 reading	Grade-4 mathematics	Grade-8 mathematics	Grade-8 mathematics
Errors in reading the test instruction scripts (additions, omissions, or revisions)	8	10	6	13
Errors in calculator distribution	n/a	5	5	10
Problems with calculator training	n/a	5	3	4
Students were disruptive or uncooperative	5	6	6	7

SOURCE: National Center for Education Statistics, *1992 Trial State Assessment: Summary of Participation Rates, Training, and Data Collection Activities* (Washington, D.C.: Author, July 1992).

Relationship between scores and characteristics of the administration. As in 1990, the Panel also considered whether certain assessment administrator behaviors, or other attributes of the assessment situation, were associated with differential student performance.[26] More competent assessment administrators (rated by Westat's quality control monitors as having greater ability to function independently overall), higher levels of student cooperation, positive student attitude toward the assessment, and higher status assessment administrators (administrators are principals, vice principals, or superintendents rather than teachers or counselors) all were consistently associated with higher levels of student performance across grades and subjects and across both years of the TSA. These associations were present even when the analyses were statistically adjusted to eliminate the effects of student gender and ethnicity on performance.

In a correlational study, it is not possible to attribute causality with certainty. Nevertheless, it seems likely that some or all of the observed associations reflect the impact of student motivation on performance. It has long been assumed that in a low-stakes assessment like NAEP, with no consequences for individual students (or even individual schools or districts), levels of student performance are somewhat dampened by low student motivation. The apparent susceptibility of performance to manipulation by the kinds of factors measured here would tend to support that hypothesis. Thus students would seem more likely to be attentive, and to perform at higher levels, when they are proctored by a higher status administrator or even simply by an administrator who is competent, organized, and capable of holding student disruption to a minimum.

Classroom observations. The Panel supplemented its analyses of quality-control-monitor data with direct observation of a small sample of sessions. Twelve schools, distributed across California, Colorado, Connecticut, Massachusetts, New Hampshire, and Rhode Island, were visited by NAE Panel staff. Although these in no way comprise a systematic sample of sessions, the investigation yielded some useful results which are outlined below.

Not surprisingly, observers noted that the mathematics assessment, which required the assessment administrators to hand out and collect various materials during the course of the assessment, was more complicated to administer and hence more subject to administrator error, than the reading assessment. In addition, certain operational repercussions were associated with lengthening the 1992 mathematics assessment to accommodate the estimation item block.[27] Students frequently became restless toward the end of the 2-hour assessment (this was especially noticeable in fourth grade), and it was more difficult to schedule the longer sessions into the class day, particularly if the school wished to use the same administrator for multiple sessions.

Student disruptions were especially noticeable during the administration of the estimation-item blocks. By pacing students through a series of problems using audiotapes, this portion of the assessment is intended to reveal whether students can provide reasonable estimates of answers without doing the actual computation. For

[26] These analyses were restricted to monitored sessions.

[27] On the average, the mathematics assessment (both grades 4 and 8) took about 2 hours, from the beginning of the script to the dismissal of students. In 1990, the grade-8 mathematics assessment averaged about 90 minutes.

1992, these blocks were always scheduled last in the assessment. In addition to general restlessness, observers noted that some students worked ahead of the tape, while others called out the answers or otherwise talked out loud during this portion of the assessment.

Some fourth-grade students were observed to have trouble answering the background questions about ethnicity, parental education, how many parents lived at home, and whether they had moved. Although NAEP uses the same wording of background questions across grades (where they overlap) and across assessment cycles, this apparently does not ensure that students in different grades attach the same meaning to their responses, and, in some cases, students may not know the relevant facts of their family history. For example, in the 1992 TSA data, about 35.7 percent of fourth-grade students and about 8.1 percent of eighth-grade students failed to provide data on parental education. Similar problems with the NAEP background variables have been noted by many researchers; Rivera and Pennock-Roman are among those who have raised questions about the validity of the NAEP race-ethnicity variable at younger ages.[28]

Summary. *In summary, although there is room for improvement, particularly with regard to the presentation of the estimation items and the collection of background data, the overall quality of the 1992 TSA administration was very high. Data available from monitored sessions were essential to reaching these conclusions. These data showed that irregularities in monitored sessions were infrequent and had limited impact on student performance. Further, the comparability of monitored and unmonitored sessions was confirmed by the fact that only very small and statistically nonsignificant differences in student performance were observed between monitored and unmonitored sessions.*

State Testing Directors' Perspectives on the Implementation of the 1992 TSA

An additional perspective on the overall implementation of the TSA was provided by the Panel's September 1992 survey of state testing directors, which paralleled a similar Panel survey carried out after the 1990 administration. Separate questionnaires were developed for the 44 participating and 12 nonparticipating jurisdictions, and written responses were received from all participating states and territories and all nonparticipating jurisdictions except one.

The responses and comments of the state testing directors reflected generally high levels of satisfaction with the assessment. All respondents from states that had participated in the 1992 TSA reported that the assessment "went well" in their states, and 80 percent gave their assessment a grade of "B" or above (98 percent gave it a grade of "C" or above). These perceptions represent a modest improvement over 1990 (when two states indicated that the assessment did not "go well" and only 75 percent gave a grade of "B" or above) and are notable in light of the greater logistical complexities occasioned by the larger scope of the 1992 assessment.

[28] C. Rivera and M. Pennock-Roman, *Issues in Race/Ethnicity Identification Procedures in the National Assessment of Educational Progress, Part I: A Comparison of Observer Reports and Self-Identification.* Research Report No. RR-87-19 (Princeton, NJ: Educational Testing Service, 1987).

Certain kinds of problems were reported with greater frequency in 1992, however. Most notably, the proportion of states reporting problems with obtaining school participation increased from 25 percent in 1990 to 48 percent in 1992, and states that felt too much staff time was required for the administration increased from 25 percent to 48 percent. Respondents also reported a higher incidence of problems with receipt of testing materials from NAEP in 1992 but fewer problems with local administrator training. Levels of problems with paperwork and arranging schedules at the school level remained about the same from 1990 to 1992, and communication problems with NAEP personnel were very infrequent in both years.

Importantly, interest in the TSA remained high or even increased between 1990 and 1992. As of the date of the survey, 89 percent of the 1992 participants were planning to sign up for the 1994 assessment, while four or five of the 1992 nonparticipants were also considering participation. The most frequently cited reasons for participating in the TSA, in 1992 as in 1990, were to have a voice in shaping state NAEP policy (64 percent), to support statewide education reforms (61 percent), and to obtain state ranking information (61 percent). Cost of participation remained an issue, however. Compared to 1990, twice as many states reported that participation in the next TSA would be contingent upon budgetary approval, and, among states that thought they were ***unlikely*** to participate, reimbursement of some or all costs by the federal government was most frequently cited as a factor that could tip the balance in favor of participation.

The Cost to the States for Participation

In 1990, the Panel conducted a small pilot study of the cost of administering the TSA. Because the data were collected retrospectively, however, there were very large variations in the amounts reported by the different states, and the data appeared unreliable. Because the costs of the TSA are a significant factor affecting state participation, the Panel conducted a more extensive study of the cost of administering the TSA in 1992.

Information about personnel time expended and resources used in the administration of the TSA were collected through written questionnaires sent to all state coordinators and a sample of 800 local assessment administrators in participating states. These questionnaires, which were distributed in advance of the assessment administration, asked respondents to keep track of state and local expenditures according to the categories shown in figure 2.5.

Figure 2.5. ***Cost categories for survey of state and local costs of administering the 1992 TSA***

STATE COSTS

State coordinator plus assistants, if any

Meetings	• Attend Network and other NAEP meetings
Recruitment	• Secure cooperation and track recruitment efforts • Review list of selected schools and update school information • Obtain assessment administrator (AA) names and shipping addresses for materials
Sampling	• Receive lists of eligible students for Westat sample selection
Coordination	• Assist with training session schedules • Schedule the assessment sessions • Send administration schedules listing selected students and principal questionnaires to AAs • Receive principal questionnaires • Send manual and training information to AAs

LOCAL COSTS

Assessment Administrators

Preparation	• Attend assessment administrator training • Receive the administration schedule and assessment materials • Update the administration schedule; select supplemental sample • Prepare, distribute, and collect assessment questionnaires • Make arrangements for the assessment
Assessment	• Prepare the room for the assessment • Conduct the assessment • Account for all materials
Conclusion	• Prepare and ship materials back to NAEP • Discuss session with NAEP staff • File the administration schedule and roster of questionnaires

Other school personnel

- Fill out Reading and Mathematics Teacher questionnaires
- Fill out School Characteristics and Policies questionnaire
- Fill out Excluded Student Questionnaires

SOURCE: C. O'Donnell, J. Chambers, and D. Ehrlich, "State and Local Costs of the 1992 Trial State Assessment," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

Usable information was received from 32 states and 2 territories. The basic unit of analysis was cost per session. This unit was chosen, rather than cost per student or per school, to separate out the effects of multiple sessions and of student absences and exclusions. Because of variation in budget and accounting procedures across states, cost estimates were generated on the basis of estimated levels of effort; they have been computed to include both in-kind and out-of-pocket costs.

As shown in table 2.10, it cost a typical state about $230,000 to participate in all three of the 1992 TSAs (fourth-grade reading, fourth-grade mathematics, and eighth-grade mathematics), but costs varied widely across states, ranging from $132,000 (North Dakota) to $372,000 (California). Because the total number of sessions also varied somewhat across states,[29] it is also instructive to consider per-session costs for participation in the TSA. *Per-session costs averaged $731, which included $50 spent at the state level and $681 spent at the local level. Thus, local costs, on average, accounted for 93 percent of all state and local costs of administering the TSA.*

[29] The average number of sessions per state was 314, but the number of sessions was substantially higher in states with high proportions of small schools and substantially lower in some very small jurisdictions.

Table 2.10. ***Cost per session and total cost, at state and local levels, combined across three assessments and across personnel and nonpersonnel costs***

State	Total state sessions	State cost per session ($)	Total state cost ($)	Local cost per session mean (SD)* ($)	Total local cost ($)	Grand total cost per session ($)	Grand total cost ($)
Alabama	314	25	7,786	577 (257)	181,142	602	188,928
Arizona	319	11	3,551	764 (480)	243,715	775	247,265
California	321	169	54,242	990 (412)	317,796	1,159	372,038
Colorado	355	47	16,857	677 (363)	240,470	725	257,327
Connecticut	319	60	19,083	895 (508)	285,499	955	304,582
Delaware	248	40	9,871	807 (392)	200,233	847	210,104
Florida	323	28	9,202	741 (369)	239,391	770	248,593
Guam	251	56	14,145	752 (798)	188,668	808	202,813
Idaho	333	40	13,405	571 (343)	190,022	611	203,427
Indiana	312	27	8,515	922 (548)	287,799	950	296,314
Iowa	362	105	38,074	518 (288)	187,370	623	225,444
Kentucky	338	6	1,958	603 (265)	203,937	609	205,895
Louisiana	322	36	11,550	532 (261)	171,437	568	182,987
Maine	285	39	11,027	469 (180)	133,656	508	144,682
Maryland	314	8	2,497	979 (800)	307,386	987	309,884
Michigan	301	21	6,280	815 (282)	245,442	836	251,722
Minnesota	306	73	22,329	576 (314)	176,389	649	198,718
Mississippi	315	15	4,789	554 (201)	174,503	569	179,292
Missouri	329	36	11,955	645 (322)	212,162	681	224,117
Nebraska	340	58	19,737	447 (219)	152,141	506	171,878
New Hampshire	301	38	11,483	555 (259)	167,114	593	178,598
New Jersey	277	136	37,737	932 (619)	258,228	1,068	295,964
New Mexico	316	24	7,634	680 (307)	214,808	704	222,442
New York	268	49	13,221	1,084 (508)	290,641	1,134	303,862
North Dakota	335	49	16,465	345 (151)	115,541	394	132,006
Oklahoma	357	41	14,493	668 (320)	238,436	708	252,929
Rhode Island	292	79	22,938	836 (550)	244,110	915	267,048
Texas	312	142	44,450	715 (299)	223,175	858	267,626
Utah	324	68	22,098	664 (383)	215,249	733	237,347
Virgin Islands	141	70	9,873	506 (576)	71,302	576	81,175
Virginia	329	29	9,387	503 (169)	165,638	532	175,025
West Virginia	384	18	7,059	646 (267)	247,978	664	255,037
Wisconsin	351	33	11,584	753 (302)	264,476	786	276,060
Wyoming	383	11	4,254	429 (181)	164,254	440	168,507
Overall Average	314	50	15,280	681 367	212,356	731	227,636
Standard Deviation	44	39	12,080	179 165	55,524	192	59,275

* Mean local level cost and standard deviation based on total sessions sampled (all three assessments).

SOURCE: C. O'Donnell, J. Chambers, and D. Ehrlich, "State and Local Costs of the 1992 Trial State Assessment," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

Additional information on per-session costs, broken out by grade level and by activity, are shown in table 2.11. As can be seen at the local level, grade-8 costs are substantially higher than grade-4 costs. This is primarily due to the fact that, at fourth grade, most schools had two assessment sessions (for reading and for mathematics), and a single assessment administrator could be responsible for both. This resulted in significant savings for session preparation, which included a half-day of training for each assessment administrator provided off-site by the NAEP contractor. Other factors affecting the differential per-session costs at grades 4 and 8 include the fact that the fourth-grade costs were averaged across the shorter reading sessions and the longer mathematics sessions (whereas only the longer mathematics sessions contributed to the eighth-grade costs), and the fact that teacher questionnaires (which had to be filled out by every mathematics [or reading] teacher who had students in the assessment) were likely to be more numerous at grade 8.

Table 2.11. ***Average state and local costs and hours per session, by activity and by grade***

Local-level average cost and hours per session, by grade				
	Grade 4		Grade 8	
	Hours per session	Cost per session	Hours per session	Cost per session
Activity				
Preparation	9.5	$283	15.0	$467
Administration	4.5	$133	5.1	$156
Conclusion	1.5	$44	2.2	$69
Questionnaire completion	2.4	$73	3.7	$115
Overall average	17.8	$540	26.0	$806

State-level average cost and hours per session, combined grade		
	Hours per session	Cost per session
Activity		
Coordination	0.7	$16
Meetings	0.3	$10
Recruitment	0.7	$18
Sampling	0.2	$6
Overall average	1.9	$50

SOURCE: C. O'Donnell, J. Chambers, and D. Ehrlich, "State and Local Costs of the 1992 Trial State Assessment," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

The most time-consuming (and therefore most expensive) activity at the state level was recruitment of schools, closely followed by coordination of training and assessment activities. Amortizing across all assessment sessions in the state, the average amounts of time spent in meetings and on sampling duties were, respectively, 0.3 or 0.2 hours per session—considerably less than the 0.7 hours per session for recruitment or coordination. However, estimates of time and cost for each of these activities were highly variable. For example, total time estimates for recruiting schools ranged from 24 hours to 560 hours, in part because some states mandated participation (thus reducing costs for recruitment), while some appeared to require considerable effort to attain their recruitment goals. Estimates for coordinating activities covered a wide range as well, from 4 hours to 880 hours. Variations in these estimates reflect, among other factors, the number of assessment sessions held in a state, the amount and cost of personnel time, and the extent of nonpersonnel costs such as travel.

Summary and Recommendations

Based on its examinations of the sampling design, evidence concerning the administration, and the data generated from the assessment results, the Panel concludes the following:

1. The TSA sample design met the standards for sample designs used in high quality, major national education surveys. It was executed with care and due attention to detail. This favorable result replicates the results of the review of the sample design conducted in 1990. However, the design did call for samples that approximated the total number of available schools, and in some cases the total number of students, in some smaller jurisdictions. Over repeated cycles of testing, this could prove problematic for states that need to repeatedly obtain cooperation from the same schools as well as test high proportions of the in-grade students at selected schools.

2. Participation rates for sampled schools have declined since the previous TSA, although they remain comparable to national NAEP participation rates. Overall, the average participation rate for initially sampled schools was about 89 percent in each of the three assessments in the 1992 TSA, compared to 86 percent for grade 4 and 88 percent for grade 8 in the national assessment. However, initial school participation rates were less than 85 percent in a third of participating states and less than 70 percent in several instances. This is a major change from 1990, when only two states had initial school participation rates below 85 percent.

3. A consistent relationship was observed between initial school participation rates and mean state scores on the TSA. Lower before-substitution participation rates were slightly associated with higher mathematics achievement in 1990, and this association was stronger in 1992. In light of the large number of states with initial participation rates below 85 percent, as well as the finding that lower initial participation rates were reliably associated with higher state scores, the Panel concludes that initial participation rates for some states were too low for accurate reporting of their 1992 TSA results.

4. The numbers of sampled students who withdrew from the schools where they were sampled before the date of the assessment substantially exceeded the numbers of students who were added from supplemental lists of newly enrolled students. This imbalance presumably reflects such factors as dropouts, delays in re-enrollment, seasonal migration to the families' countries of origin, and other temporary or permanent withdrawals from public school. However, since both dropping out and high mobility are associated with lower student performance, the bias introduced by withdrawn students who are not represented in the supplemental sample is likely to result in an overestimation of performance in states with high proportions of such students.

5. Although relatively few of the students who were absent from their original assessment sessions attended the scheduled makeup sessions, very good overall student participation rates were obtained—ranging from 93 to 97 percent for all but one state.

6. Weighting procedures used to compensate for different selection rates among different categories of students were reasonably straightforward, and adjustments for school nonparticipation were acceptable. The adjustments for student nonparticipation, however, were based on criteria that were unlikely to be very effective in removing student-level nonresponse bias. A more precise correction would require additional information on the demographic characteristics of nonparticipating students.

7. The Panel's evaluation of the excluded-student data found that rates of IEP and LEP exclusion were generally consistent between the 1990 and 1992 TSAs. Furthermore, state-by-state estimates of proportions of IEP and LEP students were in general agreement with estimates of these populations obtained from independent sources. Further study is needed to determine whether the IEP or LEP exclusion criteria were applied consistently and appropriately by local assessment administrators. Statistical sensitivity analyses carried out by the Panel suggested that the inclusion of all private school, IEP, and LEP students in the 1992 mathematics assessment would have markedly changed the mean performance and rank for a number of states.

8. Consistency in the conditions of administration for state and national NAEP is of the utmost importance in providing uniform and comparable student data. A strong test of this consistency, and, by extension, of the adequacy of the state assessment administration conditions, is to compare overall performance for each grade and subject of the TSA against the overall performance of a matched sample drawn from national NAEP. In 1992, no statistically significant differences in overall performance were observed between the TSA and matched national samples for any subject or grade.

9. The overall quality of the 1992 TSA administration was very high. Irregularities in monitored sessions were infrequent and had limited impact on student performance. Further, the comparability of monitored and unmonitored sessions was confirmed by the fact that only very small and statistically nonsignificant differences in student performance were observed between monitored and unmonitored sessions.

10. The responses and comments of state testing directors surveyed in September 1992 reflected generally high levels of satisfaction with the assessment. All respondents from states that had participated in the 1992 TSA reported that the assessment "went well" in their states, and 80 percent gave their assessment a grade of "B" or above. These perceptions represent a modest improvement over 1990—notable in light of the greater logistical complexities occasioned by the larger scope of the 1992 assessment.

    While interest in the TSA remained high between the Panel's 1990 and 1992 surveys, twice as many states reported in 1992 that participation in the next TSA would be contingent upon budgetary approval. Among states that thought they were unlikely to participate, reimbursement of some or all costs by the federal government was most frequently cited as a factor that could tip the balance in favor of participation.

11. Counting both in-kind and out-of-pocket expenses, it cost a typical state nearly $230,000 to participate in all three 1992 TSAs (fourth-grade reading, fourth-grade mathematics, and eighth-grade mathematics), but costs varied widely across states, ranging from $132,000 (North Dakota) to $372,000 (California). Per-session costs averaged $731, which included $50 spent at the state level and $681 spent at the local level. Thus, local costs, on average, accounted for 93 percent of all state and local costs of administering the TSA.

In light of these findings, the Panel makes the following recommendations for enhancing the implementation of the TSA. Additional recommendations relating to participation rates and excluded students are contained in chapters 6 and 7.

- *In order to understand the dynamics of the transfer process and create a more adequate sampling design for this segment of the population, the Panel suggests that NAEP conduct a small study to trace a subsample of students who are sampled but withdraw before the date of the assessment.*

- *NAEP should also study the feasibility of utilizing additional student-level data on absent students to improve the accuracy of adjustments for student nonresponse. Gender and race-ethnicity data, which are already collected on the administration schedules for all sampled students, would appear to be the most likely candidates.*

- *Observational data that suggest that fourth-grade students had difficulty answering a number of the student background questions, combined with the high rates of missing data for some background variables at this grade level, cast significant doubts on the usefulness of this information. Further study should be undertaken to explore methods for improving the measurement of background variables on NAEP, particularly at the lower grades.*

3 The Content Validity of the Assessment

As one might expect, appropriate decisions about test content are critical to ensuring the ultimate value and utility of any assessment. Content domains for most subjects are potentially very broad, so that the specification of an assessment design requires a whole series of thoughtful choices. In the case of classroom, or even state, assessments these choices are often made by matching test content to a prespecified instructional curriculum. This matching has the dual purpose of focusing decisions about what is most valued and appropriate in a given content area at the curriculum level, while also ensuring that students are evaluated on skills and concepts they have had the opportunity to learn.

There is no national curriculum in the United States that can serve as the basis for NAEP content specifications. We are, however, entering an era in which voluntary national curriculum standards are being established for many subject areas. Specifications for recent NAEP assessments have been designed to contribute to the national dialogue on standards by aligning both the content and the form of the assessment with new visions of education and literacy. At the same time, recent NAEP consensus projects have grappled with some competing priorities for content specification, particularly the extent to which the content for a state-by-state NAEP should be broadly representative of current curricula in the states. The Panel, in its first report to Congress, concluded that NAEP could best fulfill its key role as an independent monitor of educational progress by providing *comprehensive* coverage of content and assessment methodologies, "creating an appropriate balance between established practice and cutting-edge advances in education."[1]

Objectives and Frameworks for the 1992 Trial State Assessment

The same law that gave rise to the trial state assessments directed that NAEP content objectives be developed through a consensus process involving the "active participation of teachers, curriculum specialists, subject-matter specialists, local school administrators, parents, and members of the general public."[2] The consensus efforts for each of the two subjects assessed in the 1992 TSA—mathematics and reading—were contracted to CCSSO, which assembled, in each case, a steering committee and a planning committee to assist in its work. In addition, departments of education in each of the 50 states were brought into the specifications process to ensure that there would be an acceptable level of alignment between NAEP content and state curricula.

[1] The National Academy of Education, *Assessing Student Achievement in the States* (Stanford, CA: Author, 1992), 29.

[2] Public Law 100-297, Part C, Section 3403 6A: 1988.

The mathematics objectives were the first to be created under this new model. They were developed in 1988 and used in the 1990 and 1992 assessments. Original plans called for the same objectives or "framework" to be continued through additional assessment cycles to facilitate the measurement of trends in student performance. However, the rapid evolution of contemporary perspectives on mathematics education and assessment led to a rethinking of this position, with the result that the mathematics objectives were rewritten in anticipation of the next assessment cycle—originally expected to occur in 1994, but now likely to be delayed until 1996.

The reading objectives were the next to be written, in 1990. These have been followed by frameworks for science, geography, U.S. history, and art, each of which may someday become part of the NAEP state assessment program.[3] The objectives-development process thus has been an evolving one, with each consensus project building on the lessons of the last. In fact, each of the projects has been required to submit a document reflecting on its processes and including recommendations for future consensus projects. Building on the recommendations of the mathematics project, for example, all subsequent consensus projects (including reading) have been required to produce detailed test specifications in addition to the more public, and more philosophical, objectives/framework document. And, with the emergence of the achievement levels as a reporting device, consensus projects operating from 1992 forward have also become responsible for articulating subject- and grade-specific definitions for these levels.

Further, in keeping with contemporary thinking on assessment, recent consensus projects have considered the form, as well as the content, of the assessments. Unlike earlier models of learning and assessment, in which it was assumed that test format was a purely technical decision to be made by psychometricians once the desired content had been specified by subject-matter experts, the current paradigm holds that the form of the assessment is inextricably intertwined with the content. In particular, it is widely believed that many of the most important educational attainments require complex, integrated behaviors that cannot be successfully disaggregated during teaching or testing. The consensus projects, therefore, have taken it under their purview to specify the nature, and even the proportion, of performance-type tasks to be included in the assessments. While this aspect of assessment design was least articulated in the mathematics project that led to the 1990 objectives, the more detailed test specifications that were developed by each subsequent project have provided a better vehicle for such efforts.

Mathematics Framework

The objectives for the 1990 and 1992 mathematics assessments, along with the consensus process that gave rise to these objectives, were reviewed by the Panel in its first report. In summary, the Panel concluded that

> a national consensus process is an appropriate basis for defining NAEP content objectives and frameworks. The consensus process that generated the 1990 mathematics objectives and framework was

[3] Plans are also underway to develop a Civics Framework, although the project has been delayed by budget considerations.

> judged adequate, although not as broadly representative as would have been desirable. ... The 1990 NAEP Mathematics Framework and the assessment design principles on which it was built were essentially sound. Further, the framework represented a reasonable compromise between current instructional practices and the standards being put forward by the professional mathematics community at that time. To remain viable, the objectives and frameworks must be viewed as evolutionary entities which are adapted regularly to reflect changing conceptions of mathematics education and measurement.[4]

The last point is particularly important. Shortly after the objectives were completed in 1988, the *Curriculum and Evaluation Standards for School Mathematics* were finalized by NCTM and widely endorsed as a significant statement concerning the ways in which mathematics should be taught in the schools.[5] Although the NAEP objectives had been influenced by draft versions of the standards, the wide and rapid acceptance subsequently attained by the standards quickly shifted the mathematics community's perceptions of what should be included in a national assessment. In effect, the 1990 NAEP was beginning to appear somewhat outdated before it occurred. This dissonance became even more noticeable as the 1992 assessment approached, and it led to NAGB's decision to implement some modest changes in the makeup of the 1992 assessment (as will be discussed below) as well as to update the framework itself before using it in any further assessment cycles.

To what extent do the changes to the 1992 assessment and the subsequent revisions of the framework represent a reasonable response to the rapid evolution of mathematics education? There is no clear answer. On the one hand, a NAEP that falls out of step with national content standards loses its utility and credibility, and the professional mathematics community, as well as this Panel,[6] strongly urged that NAEP improve its alignment with the NCTM Standards. On the other hand, NAEP is charged with the mission of measuring progress over time, and this mission becomes substantially more problematic when the measurement tool itself is subject to a process of rapid change. In its recent report on setting performance standards for student achievement, the Panel suggested that an 8- to 10-year implementation, feedback, and revision cycle is appropriate. "Stability over some sustained period of time is important substantively, technically, and educationally. An 8- to 10-year cycle is long enough to establish meaningful trend lines, but short enough to accommodate the natural evolution that occurs in subject-matter fields."[7]

[4] The National Academy of Education, op. cit., 19-20. Further details on the 1990 Mathematics Framework and consensus process can be found in E.A. Silver, P.A. Kenney, and L. Salmon-Cox, "The Content and Curricular Validity of the 1990 NAEP Mathematics Items: A Retrospective Analysis," in *Assessing Student Achievement in the States: Background Studies* (Stanford, CA: The National Academy of Education, 1992).

[5] National Council of Teachers of Mathematics, *Curriculum and Evaluation Standards for School Mathematics* (Reston, VA: Author, 1989).

[6] The National Academy of Education, op. cit., 29.

[7] The National Academy of Education, *Setting Performance Standards for Student Achievement* (Stanford, CA: Author, 1993), 112.

After considering the alternatives, the Panel concludes that the present rate of change in the NAEP mathematics content is reasonable. However, as the assessment continues to evolve, it will be important to investigate empirically the question of whether the score scale continues to have the same meaning over consecutive administrations. In the long run, NAEP may need to reconceptualize its reporting practices in order to adequately capture student attainment against both traditional and emerging content standards. For example, as is suggested in chapter 8 of this report, it may prove more feasible to move to a reporting structure that provides two separate scores: one for the more traditional content and one for the new content.

Reading Framework

The context for the development of the 1992 Reading Framework was different in a number of ways. On the positive side, the reading project was the second consensus project to develop content specifications for a state-by-state NAEP. As such, it could benefit from the experiences and conclusions of the first effort (in mathematics). On the negative side, the reading project, unlike the mathematics project, could *not* draw on a set of national content standards endorsed by the full professional community. Rather, the professional reading community was characterized by "some of the deepest and longest running schisms in education," affecting concepts of the nature of reading as well as views of how reading should be taught.[8] Further, one of the largest and most influential professional associations in reading, the International Reading Association (IRA), refused to support the development of the TSA in reading based on an existing organizational resolution opposing the "proliferation of inappropriate assessments" and comparisons of states, districts, or schools based on "inappropriate assessments."[9]

Consensus project. Like the mathematics project, the reading consensus project was guided by a steering committee of policymakers, practitioners, and citizens who established basic principles and policies. Some of these principles were designed to help the project sidestep the factional disputes within the reading community. In particular, the steering committee mandated a framework that would focus on outcomes or performance, rather than any particular instructional or theoretical approach, and one that would account for contemporary research on reading and literacy without, however, departing from the primary focus on outcomes.[10] Reading project members felt that these strategies did, in fact, prove to be successful. In their final report, they recommended to future consensus projects that

> Scientifically sound research on the learning of a subject or a process like reading can generally be reconciled or resolved into an inclusive model that is also sound and useful. When research on a subject reflects different or competing models, endeavoring to synthesize the

[8] Council of Chief State School Officers, *Report of the Consensus Process: 1992 NAEP Reading Assessment National Consensus Planning Project* (Washington, D.C.: Author, 1990), 6.

[9] International Reading Association, *Assessment Resolutions* (Newark, DE: Author, 1991).

[10] NAEP Reading Consensus Project, *Reading Framework for the 1992 National Assessment of Educational Progress* (Washington, D.C.: National Assessment Governing Board, 1992), 5.

> traditions and applying the principle of accounting for sound findings appear to be successful consensus planning strategies.[11]

And,

> In subjects where professional differences exist about the teaching of the subject, assessment recommendations can be aimed at mutually-agreed desirable ends of instruction, rather than the means toward those ends. This is especially workable in National Assessment, where testing is done at three major stages of schooling and not at many grade levels throughout the process.[12]

Other principles developed by the steering committee underscored NAGB's charge to the consensus effort to "range fully and think openly and widely about the assessment" and not be "constrained or compromised by past practice or current resource constraints."[13] These principles included a directive to "expand the range of assessment tools to include such techniques as open-ended questions and other new approaches and formats appropriate to the objectives," as well as the aforementioned emphasis on the use of research findings.[14]

The 15-person planning committee, which had technical responsibility for drafting the framework and specifications, was composed primarily of experts in reading and included college professors, a university dean, a classroom teacher, a local school administrator, state-level specialists in reading and assessment, and a representative of the business community. Throughout the project, the perspectives of these committee members were supplemented by inputs and reactions from members of the reading field representing a wide range of perspectives, as well as experts in assessment, school administrators, and state staff in reading and assessment. In addition, project staff included a full-time coordinator who was also a member of the professional reading community and a subcontractor with considerable experience in the development of test specifications.

This level of staff support was an improvement over the mathematics project and was responsive to that project's recommendations. The expansion of the consensus effort to include the development of detailed test specifications was another well-received enhancement that was informed by the suggestions of the mathematics project. Regrettably, however, in contracting for the reading project, *NAGB was not able to heed what was perhaps the most impassioned criticism delivered by the mathematics project—the plea for more time.*[15] *In fact, the reading project was worse off, having*

[11] Council of Chief State School Officers, *Report of Consensus Process*, 8.

[12] Ibid., 9.

[13] Ibid., 13-14.

[14] NAEP Reading Consensus Project, *Reading Framework*, 5.

[15] Council of Chief State School Officers, *Report of the Consensus Planning Process: 1990 NAEP Mathematics Assessment* (Washington, D.C.: Author, 1988).

only 6 months—1 month less than the mathematics project—to accomplish, in a more fractious environment, what was now (with the inclusion of the test specifications) a larger body of work.[16]

As noted by the experts that the Panel commissioned to review the content validity of the reading assessment (as well as by members of the reading project itself), this extremely tight timeline resulted in certain negative consequences. It strained the personal resources of the participants, left rough edges and troubling inconsistencies in the documents prepared for public dissemination, and most critically, allowed little time for meaningful discourse and consensus building with the wider professional community.[17]

Further, the short timeline precluded the project from undertaking any significant advances in measurement methodology. Consensus-project participants interviewed by the Panel's reading experts commended the cross-discipline cooperation evident on the project; one participant, for example, was quoted as saying, "I think that there was every effort made on the part of the measurement community to listen to the people in reading, and I think that if we had given better constructs they would have done things to support that. I feel that they have responded always with a great deal of respect to the people in the literacy community and been very responsive."[18] But it would be unrealistic to expect much in the way of true methodological innovations from a project bounded by such severe time constraints. Rather, the project searched for, and, where possible incorporated, innovative measurement strategies culled from the more forward-looking state assessments and the research literature in reading.

Despite these constraints, the Panel concludes that the reading consensus project produced a framework that represents a substantial advance over previous NAEP Reading Frameworks and is reasonably responsive to most of the current theories and practices in reading.

The framework. The 1992 Reading Framework and specifications included a number of significant innovations:

- a definition of a "good reader" based upon a constructivist view of learning;
- use of authentic texts;
- assessment of reading across different reading situations;
- attention to the various cognitive aspects of reading;
- a high proportion of open-ended assessment items; and
- more time for students to respond to each reading passage.

[16] The entire consensus project lasted for 12 months, but the essential structure of the framework and the specifications had to be completed within 6 months to allow item development to begin.

[17] B.C. Bruce, J. Osborn, and M. Commeyras, "The Content and Curricular Validity of the 1992 National Assessment of Educational Progress Reading Framework," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

[18] Ibid.

The framework classified items according to reading situation and according to the type of interaction involved between the reader and the text (referred to as "aspects of reading"). Reading situations were classified as reading for literary experience, reading to be informed, and reading to perform a task. Aspects of reading were classified as forming an initial understanding, developing an interpretation, personal reflection and response, and demonstrating a critical stance. These aspects of reading are further explicated in figure 3.1.

Figure 3.1. ***Framework for the 1992 reading assessment: Aspects of reading***

Forming an initial understanding	Developing an interpretation	Personal reflection and response	Demonstrating a critical stance
Requires the reader to provide an initial impression or unreflected understanding of what was read.	Requires the reader to go beyond the initial impression to develop a more complete understanding of what was read.	Requires the reader to connect knowledge from the text with his/her own personal background knowledge. The focus here is on how the text relates to personal knowledge.	Requires the reader to stand apart from the text and consider it.

SOURCE: NAEP Reading Consensus Project, *Reading Framework for the 1992 National Assessment of Educational Progress* (Washington, D.C.: National Assessment Governing Board, 1992), 16-17.

The reading planning committee specified that the assessment be constructed of reading passages drawn from authentic texts used by students in real, everyday reading. Whole stories, articles, or sections of textbooks should be used rather than excerpts or abridgements. "Passages written solely for a test or to provide drill in a specific skill were not considered appropriate for this assessment."[19] A high percentage of assessment items were to be open ended, including both short-answer items and longer, extended-response items that would be scored using a partial-credit model. In addition, the committee asked for longer assessment blocks than had been used in recent NAEP assessments to allow time for students to read the somewhat longer authentic passages and respond meaningfully to the new constructed-response items.[20]

Finally, the 1992 Reading Framework provided a context for the assessment by expanding on the meaning of reading literacy and the goals of a literary education. The term "reading literacy," as used by the framework, "is not intended to imply only basic or functional literacy. Rather the term connotes a broader sense of reading, including knowing when to read, how to read, and how to reflect on what is read."[21] The goal of literacy education is to develop good readers, that is, readers who

[19] NAEP Reading Consensus Project, *Reading Framework*, 19.

[20] An item block is a set of items that always appear together. A student must answer all of the items in an item block within a single timed period.

[21] NAEP Reading Consensus Project, *Reading Framework*, 6.

> (a) possess positive habits and attitudes about reading; (b) read with enough fluency so that they can focus on the meaning of what they read; (c) use what they already know to understand what they read; (d) form an understanding of what they read and extend, elaborate, and critically judge its meaning; (e) use a variety of effective strategies to aid their understanding and to plan, manage, and check the progress of their reading; and (f) can read a wide variety of texts and can read for different purposes.[22]

The authors of the framework understood that not all of these characteristics of good readers could be assessed, using existing methodologies, in a project of national scope. However, they cautioned that

> the framework must not promote assessment items that in terms of correlating with reading ability, have only statistical utility, but do not correspond to sound concepts of reading or resemble desirable classroom activities. ... every effort must be exerted to make the best use of available methodology and resources ... [and] capabilities in assessment must be driven forward, so that important aspects of reading that now elude measurement can be addressed in the near future.[23]

Review by experts. A sample of 21 diverse and respected reading experts interviewed shortly after the release of the 1992 reading results evaluated the framework in light of its intended purpose as a basis for a national assessment. *On average, these experts were reasonably satisfied with the framework. Most gave it a "B" grade, two gave it an "A," and three gave it a "C" (table 3.1).*

Table 3.1. ***Expert evaluation of the 1992 Reading Framework—based on October 1993 survey (total n = 21)***

	A	B+	B	B-	C+	C	C-
Overall, what grade would you give to the framework, given the purpose for which it is intended?	2	3	12	1	1	1	1

SOURCE: J. Mitchell, "Evaluation of the 1992 Reading Framework," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

[22] Ibid., 9.

[23] Ibid., 7.

In general, the reading experts commended the framework for those same characteristics that set it apart from previous NAEP Reading Frameworks, starting with the constructivist definition of reading that takes account of the effects of reader, text, and context on the construction of meaning. The experts also praised the attention to different types of text and reading purposes, the movement away from multiple-choice response modes and a skills-based definition of reading, the use of longer and authentic passages, the attention to multiple factors in judging difficulty of text, the sound theoretical basis of the framework, and the relevance to everyday classroom practices.

Several of the reading experts expressed concern, however, that the conceptualization of reading articulated in the framework, while highly desirable, could not be adequately operationalized in a paper-and-pencil, on-demand assessment such as NAEP. For example, while the framework proposes to assess reading in a variety of situations and acknowledges that meaning is influenced by context (e.g., reading a train schedule in a train station, surrounded by clues that contribute to the interpretation of the document), the test taker can actually be reading for only one purpose—to complete the assessment. Similarly, some of the experts noted that although the integration of reading and writing through the considerable use of constructed-response items was a major step forward for the framework, the assessment still fell short of measuring important social aspects of reading and the integration of reading (and writing) with speaking and listening. A few other experts, however, felt that the framework had moved *too far* ahead of current practice. They expressed concerns, for example, that the heavy reliance on constructed-response items would lead to an underestimation of the reading abilities of students who lacked writing skills and that the emphasis on higher order outcomes might result in an assessment incapable of measuring more basic levels of literacy.

Finally, a number of the reading experts found the sample items and scoring rubrics included in the framework document to be less than exemplary. These samples had been completed in haste to meet the demanding schedule of the framework-development process, as discussed above. The samples will be replaced with more carefully crafted materials when the framework is reissued for the 1994 assessment cycle.

Implementation of the Frameworks—Items for the 1992 TSA

Frameworks provide the template for an assessment design. Ultimately, however, it is the manner in which the frameworks are translated into actual assessment items that determines the content validity of the assessment.

Mathematics Items

The framework for the 1992 mathematics assessment, unchanged from the 1990 assessment, was a five-by-three matrix of content areas and mathematical processes. Separate subscales were computed and reported for the five content areas but not for the three types of mathematical processes. Tables 3.2 and 3.3 show how the assessment items were to be distributed over the two matrix dimensions.

Table 3.2. ***Framework for the 1992 mathematics assessment: Percentage distribution of items by grade and content area***

	Numbers and operations	Measurement	Geometry	Data analysis, statistics, and probability	Algebra and functions
Grade 4	45	20	15	10	10
Grade 8	30	15	20	15	20
Grade 12	25	15	20	15	25

SOURCE: Educational Testing Service, *Mathematics Objectives: 1990 Assessment* (Princeton, NJ: Author, 1988), 14.

Table 3.3. ***Framework for the 1992 mathematics assessment: Percentage distribution of items by grade and mathematical process***

	Conceptual understanding	Procedural knowledge	Problem solving
Grade 4	40	30	30
Grade 8	40	30	30
Grade 12	40	30	30

SOURCE: Educational Testing Service, *Mathematics Objectives: 1990 Assessment* (Princeton, NJ: Author, 1988), 14.

Despite the fact that the framework stayed the same between 1990 and 1992, there were a number of changes made to the 1992 item pool. For one thing, the 1992 assessment contained more items overall.[24] The numbers of difficult items and open-ended items were also increased, and two new item types were added for the first time in 1992. These new item types were extended-response open-ended items and items using manipulatives. (The former "allowed students to demonstrate—in writing

[24] Due to the matrix-sampling design of the national assessment, the number of items attempted by any one student remained unaffected by the overall increase in items.

and diagrams—their mathematical reasoning and problem-solving ability,"[25] and the latter provided the students with cardboard geometric shapes that they could manipulate to help them reason through a problem.) All of the changes were in the direction recommended by the mathematics experts who evaluated the content validity of the 1990 TSA on behalf of the Panel.[26]

Distribution of items. Excluding the special study estimation items, the 1992 TSA for grade 4 included 158 items, 45 percent more than the pool of nonexperimental fourth-grade items in 1990. Forty-one percent of these items were open ended, which represented a 60 percent increase over 1990. The proportion of fourth-grade items in which students were allowed to use a calculator remained about constant, and each of the two new item types was included, although in small numbers (five extended-response items and six manipulative items). Surprisingly, the content and process distribution of fourth-grade items, as classified by the NAEP contractor, deviated somewhat from the distribution specified in the framework. Specifically, there were only two-thirds as many procedural knowledge items as specified, while the proportion of problem-solving items was increased by about one quarter. Compared to the framework, there was also a slight underrepresentation of items from the numbers and operations content area, with small corresponding overages for each of the other content strands. After interviews with the test developers, the mathematics experts commissioned by the Panel concluded that these deviations from the framework were not by deliberate intent but were the result of giving priority to other important item attributes during the item-selection process.[27]

Findings for the eighth-grade items were similar. There were 34 percent more eighth-grade items overall, and the proportion of open-ended items increased from 26 percent to 41 percent. Six of the items required an extended response, and seven used manipulatives. Compared with the content and process distribution specified in the framework, the eighth-grade item pool included fewer algebra and functions items than expected (20 percent intended versus 15 percent actual) and substantially more problem-solving items (30 percent intended versus 40 percent actual).

Fit to the framework. As in the evaluation of the 1990 TSA, the Panel's study of mathematics content validity included an independent analysis of the fit of the items to the NAEP framework (as judged by the level of agreement between the official NAEP item classifications and the classifications made by the Panel's mathematics experts) as well as an analysis of the fit to the NCTM Standards. A group of seven mathematics experts were assembled for these analyses, which were restricted to the fourth-grade item pool. The experts generally worked in pairs, with some items being classified by all pairs to provide a measure of interrater agreement. For classification by the framework content categories, interrater agreement was 93 percent. As expected based on the 1990 experience, the interrater agreement rate was lower for

[25] I.V.S. Mullis, J.A. Dossey, E.H. Owen, and G.W. Phillips, *NAEP 1992 Mathematics Report Card for the Nation and the States* (Washington, D.C.: National Center for Education Statistics, April 1993), 37.

[26] The report of the Panel's evaluation of the 1990 TSA was not released until after the 1992 assessment had been administered and thus, obviously, could not have directly influenced that assessment. However, the concordance between the recommendations and the changes is not surprising given that the Panel's content recommendations reflected the general sentiments of the professional mathematics community.

[27] E.A. Silver and P.A. Kenney, "The Content and Curricular Validity of the 1992 NAEP TSA in Mathematics," in *The Trial State Assessment: Prospects and Realities. Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

the classifications by the framework process categories (73 percent). This lower level of agreement was attributed to the fact that the processes are difficult to apply (or assess) in isolation, and many of the assessment items actually draw on two or more of the process categories. Disagreements with regard to process most often involved conceptual understanding versus either procedural knowledge or problem solving.

As shown in tables 3.4 and 3.5, agreements between the consensus of the Panel's mathematics experts and the official NAEP classifications were also higher for the content classifications and lower for the process classifications. Agreement by content classification was 89 percent, while agreement by process classification was 69 percent. Again, the greatest numbers of disagreements were in the area of conceptual understanding, where agreement was only 58 percent. *In general, the disagreements highlight the shortcomings of the matrix design used for the 1992 Mathematics Framework, which required that each item be associated with one and only one framework cell. The new 1994 framework addresses this problem by discontinuing the matrix framework in favor of a model consisting primarily of the five major content strands.*[28]

Table 3.4. ***Content classification of 1992 fourth-grade TSA mathematics items: Cross-classified by NAEP and by expert raters***

NAEP	Expert raters					Total items as judged by NAEP
	Numbers and operations	Measurement	Geometry	Data analysis, statistics, and probability	Algebra and functions	
Numbers and operations	53 (84%)	6	—	—	4	63
Measurement	—	28 (93%)	1	—	1	30
Geometry	—	2	24 (92%)	—	—	26
Data analysis, statistics and probability	—	—	—	20 (100%)	—	20
Algebra and functions	2	—	—	1	14 (82%)	17
Total items as judged by expert raters	55	36	25	21	19	156

NOTE: Diagonal (percent agreement) entries based on NAEP (row) totals. Two items were not included in the cross-classification because the expert raters could not agree on a classification.

SOURCE: E.A. Silver and P.A. Kenney, "The Content and Curricular Validity of the 1992 NAEP TSA in Mathematics," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

[28] National Assessment of Educational Progress, *1994 National Assessment of Educational Progress Mathematics Assessment: Framework and Specifications* (Princeton, NJ: Educational Testing Service, March 1992), unpublished report. Note that the framework was written for use in the 1994 assessment. However, it now appears that mathematics will not be assessed again by NAEP until 1996.

Table 3.5. ***Process classification of 1992 fourth-grade TSA mathematics items: Cross-classified by NAEP and by expert raters***

NAEP	Expert raters			Total items as judged by NAEP
	Conceptual understanding	Procedural knowledge	Problem solving	
Conceptual understanding	38 (58%)	15	13	66
Procedural knowledge	1	25 (78%)	6	32
Problem solving	6	7	46 (78%)	59
Total items as judged by expert raters	45	47	65	157

NOTE: Diagonal (percent agreement) entries based on NAEP (row) totals. One item was not included in the cross-classification because the expert raters could not agree on a classification.

SOURCE: E.A. Silver and P.A. Kenney, "The Content and Curricular Validity of the 1992 NAEP TSA in Mathematics," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

Match to NCTM Standards. As was discussed earlier in this chapter, the NCTM Standards were still in draft form at the time that the 1990-1992 NAEP Mathematics Framework was written, and the framework was not intended as an exact embodiment of the standards. Since then, however, the professional mathematics community has urged NAEP to improve its alignment with the standards, and NAGB has expressed agreement with this goal. *In the judgment of the Panel's mathematics experts, nearly 60 percent of the 1992 fourth-grade items were related in some nontrivial way to one or more of the NCTM Standards themes. This represents an improvement over the 199X grade-8 item pool, in which only about 50 percent of the items had been judged by a similar panel to match any of the major themes of the standards.*[29]

Table 3.6 shows the distribution of 1992 fourth-grade items according to the standards themes they fit best. As can be seen, the themes of problem solving and reasoning were most heavily represented, while there were relatively few items that primarily related to the themes of communication or connections. The ratio of multiple-choice to

[29] E.A. Silver, P.A. Kenney, and L. Salmon-Cox, "The Content and Curricular Validity of the 1990 NAEP Mathematics Items: A Retrospective Analysis," in *Assessing Student Achievement in the States: Background Studies* (Stanford, CA: The National Academy of Education, 1992).

nonmultiple-choice items among standards-related items was similar to the ratio found in the full item pool. However, among the 10 items classified as communication, 8 were nonmultiple-choice. This suggests that open-ended or extended-response item formats are an appropriate way to assess mathematical communication.

Table 3.6. ***Distribution of 1992 fourth-grade items according to themes from the NCTM Standards***

Problem solving	Reasoning	Communication	Connections	None
37 (23%)	39 (25%)	10 (6%)	8 (5%)	64 (41%)

SOURCE: E.A. Silver and P.A. Kenney, "The Content and Curricular Validity of the 1992 NAEP TSA in Mathematics," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

Although nearly 60 percent of the items matched the standards in some nontrivial way, a much smaller percentage of the items (9 percent) were judged to be exemplars of the standards themes. With respect to item formats, six exemplary items were multiple choice, five were open ended, and three were extended response. Compared to the total item pool, nonmultiple-choice items were thus overrepresented among the items chosen as exemplary.

Special item types. The Panel's mathematics experts, although they had a number of minor concerns and suggestions for improvement, generally reacted favorably to the new extended-response and manipulative items included in the grade-4 assessment, noting that most of these items were set in an interesting context and assessed mathematically important ideas. *A serious problem with the extended-response items, however, was that substantial numbers of students skipped these problems without attempting to answer them.* For those that did answer them, insufficient time may have been available for a fully developed response. The mathematics experts felt that the placement of extended-response items at the end of item blocks exacerbated both of these problems.

Another area of criticism concerned the items designed to be solved with the aid of a calculator. As was the case with the 1990 grade-8 item pool, the mathematics experts concluded that the majority of these items involved only uninteresting and unnecessarily tedious calculations. The experts also recommended that students be allowed to use calculators on most or all item blocks, rather than only a few special item blocks. This would be consistent with the NCTM Standards view that "calculators should be available for all students at all times."[30] In this connection, the 1994

[30] National Council of Teachers of Mathematics, op. cit., 8.

framework represents an advance, as it directs NAEP to investigate the unrestricted use of calculators on all but trend blocks or specifically excluded items.[31]

Other item-design issues. In considering ways in which the overall quality of the fourth-grade mathematics items might be improved, the mathematics experts noted that a number of items used words or concepts that typical fourth-grade students may not have had an opportunity to learn; examples include the use of the word "diagonal" and the use of pie charts and line graphs. The experts also pointed out that some items were worded in ways that could be confusing, especially for students with weak reading skills. At the same time, they felt that some of the items designed to evoke an extended response failed to use language in ways that would encourage students to take a mathematical approach to the problem or to use appropriate mathematical terminology in their responses.

Item difficulty. In a separate study undertaken by the Panel to explore alternative methods for setting achievement levels, items were mapped onto the NAEP scale (by placing each item at the score point at which it would be answered correctly by 80 percent of examinees achieving that score, without guessing) and were then reviewed by mathematics experts.[32] This allowed the experts to observe patterns of item performance and to identify gaps in which there were few indicators of student achievement. A paucity of items that were not only difficult, but also measured higher level content was made clearly evident by this process, despite the fact that the NAEP contractor had attempted to strengthen the upper range of the item pool between 1990 and 1992.

In summary, the Panel concludes that the items for the 1992 mathematics assessment represented an advance over the 1990 item pool with broader coverage of the content domain, better fit to the NCTM Standards, and the introduction of interesting new item formats. However, the assessment was still inadequate for measuring advanced achievement on challenging subject matter.

Reading Items

As discussed above, the framework for the 1992 reading assessment classified items according to reading situation and reading aspect. Separate subscales were computed, and are scheduled to be reported, for the three reading situations.

The framework calls for items to be distributed across reading situations according to the proportions shown in table 3.7. The distribution was intended to be roughly proportional to the distribution of reading situations encountered at each grade level. Note that at grade 4, the assessment does not include reading to perform a task. The

[31] National Assessment of Educational Progress, op. cit. A trend block is a block of items that remains unchanged from assessment to assessment in order to facilitate the measurement of trends in student performance.

[32] E.A. Silver and P.A. Kenney, "Expert Panel Review of the 1992 NAEP Mathematics Achievement Levels," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993).

consensus project arrived at this solution because they concluded that reading to perform a task accounted for only a small fraction of total reading time at grade 4 and that, if these items were to be represented proportionately, there would be insufficient items to form a subscale.

Table 3.7. ***Framework for the 1992 reading assessment: Percentage distribution of items by grade and reading situation***

	Literary experience	To be informed	To perform a task
Grade 4	55	45	0
Grade 8	40	40	20
Grade 12	35	45	20

SOURCE: NAEP Reading Consensus Project, *Reading Framework for the 1992 National Assessment of Educational Progress* (Washington, D.C.: National Assessment Governing Board, 1992), 14.

The distribution of items across aspects of reading was laid out in terms of testing time rather than numbers of items. According to the reading specifications, one third of the testing time should be devoted to personal reflection and response, one third should be devoted to demonstrating a critical stance, and the remaining one third of the testing time should be divided between forming an initial understanding and developing an interpretation. The specifications also directed the test developer to prepare an assessment in which 40 percent of the testing time was devoted to extended-response and other open-ended items.

Fit to the framework. In order to evaluate the fit to the framework, the Panel convened a group of 12 reading experts to analyze the reading items much as the mathematics experts had analyzed items from the mathematics assessment. In the case of reading, items at all three grade levels were reviewed, with four experts working together at each grade level.[33]

Reading situation, as operationalized on the 1992 NAEP, was closely aligned with the type of text (e.g., fiction, nonfiction, and documents). This was of some concern to the expert panelists, who felt that it obscured the fact that literature can be read to inform and nonfiction can provide pleasure as well as information. In their own classifications, the experts moved some of the items associated with fictional passages into the "reading for information" category and some of the items associated with nonfiction into the "reading for literary experience" category. At grade 12, they also classified one set of items with an official NAEP designation of "reading to perform a task" into "reading for understanding."

[33] D. Pearson and L. DeStefano, "Content Validation of the 1992 NAEP in Reading: Classifying Items According to the Reading Framework," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

Overall, the classification by the expert raters suggested that the actual assessment contained a somewhat larger percentage of items tapping reading for literary experience than the framework specified. This is summarized in table 3.8, which can be compared to the framework specifications in table 3.7. At each grade, there were an additional 5 percent to 8 percent of items in the literary experience category, with corresponding decreases in the other reading situations. *Despite these perturbations, however, the reading experts considered the distribution of items across reading situations to be a reasonable approximation of the ideal described in the Reading Framework.*

Table 3.8. ***Independent expert classification of items from the 1992 reading assessment: Percentage distribution by grade and reading situation***

	Literary experience	To be informed	To perform a task
Grade 4	60	40	0
Grade 8	47	41	12
Grade 12	43	40	17

SOURCE: D. Pearson and L. DeStefano, "Content Validation of the 1992 NAEP in Reading: Classifying Items According to the Reading Framework," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

Disagreements between the Panel's reading experts and the official NAEP classifications were more frequent for aspects of reading. To some extent, these disagreements reflected the fact that aspects of reading, like the mathematical processes in the Mathematics Framework, are not cognitively discrete. In most instances, responding to an assessment item involves two or more of the aspects. However, other substantive concerns also contributed to the disagreements. For example, some of the experts were critical of scoring procedures for constructed-response items that appeared to assess critical stance, yet required students to give examples from the text. Citing the concern that such scoring requirements run the risk of discouraging students from "standing apart" from the text, the experts tended to classify these items as developing interpretation rather than critical stance. This resulted in a distribution of items (by testing time) that was substantially skewed toward "initial understanding" and "developing interpretation" compared to the specifications (see table 3.9).[34] The experts judged critical-stance items in particular to be strikingly absent from the assessment at grades 4 and 8.

[34] Testing time was computed based on the following information from ETS: 1 minute per short-answer constructed-response item, 3 minutes per extended constructed-response item, 55 seconds per multiple-choice item at grade 4, and 45 seconds per multiple-choice item at grades 8 and 12.

Table 3.9. ***Independent expert classification of items from the 1992 reading assessment: Percentage distribution of assessment time by grade and aspect of reading***

	Forming an initial understanding/ developing an interpretation	Personal reflection and response	Demonstrating a critical stance
Required by specifications	33	33	33
Expert classification Grade 4	67	24	9
Expert classification Grade 8	86	11	3
Expert classification Grade 12	59	12	29

SOURCE: D. Pearson and L. DeStefano, "Content Validation of the 1992 NAEP in Reading: Classifying Items According to the Reading Framework," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

The areas of disagreement can be seen more clearly in the cross classification of aspects of reading by NAEP and by the expert raters shown in table 3.10. In reviewing this table, it is important to bear in mind that the official NAEP distributions by numbers (or percentages) of items and by testing time are not identical. Rather, these two types of official classifications diverge because the extended constructed-response items, which have longer response times, are concentrated in the personal-response and critical-stance categories.

Table 3.10. ***Aspect classification of 1992 TSA reading items (all grades combined): Cross-classified by NAEP and by expert raters***

NAEP	Expert raters				Total items as judged by NAEP
	Forming an initial understanding	Developing an interpretation	Personal reflection and response	Demonstrating a critical stance	
Forming an initial understanding	12 (50%)	11	1	0	24
Developing an interpretation	46	97 (66%)	1	3	147
Personal response and reflection	3	9	33 (64%)	7	52
Demonstrating a critical stance	16	61	13	28 (24%)	118
Total items as judged by expert raters	77	178	48	38	341

NOTE: Diagonal (percent agreement) entries based on NAEP (row) totals. Does not include items from the special NAEP reader blocks at grades 8 and 12.

SOURCE: D. Pearson and L. DeStefano, "Content Validation of the 1992 NAEP in Reading: Classifying Items According to the Reading Framework," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

Open-ended items. The specifications called for 40 percent of the assessment time in each grade to be devoted to open-ended items. However, materials released by ETS indicate that the actual amount of assessment time allocated for open-ended items was substantially higher—varying from a low of 61 percent for grade 4 to a high of 74 percent for grade 12 (table 3.11).[35] This increased emphasis on open-ended items is a reflection of the speed with which NAEP is moving to align its assessments with current concepts in the subject-area professional communities. At the time the framework and specifications were written (1990), the planning committee favored an assessment that was primarily constructed response, but the committee was concerned that such a design might not be feasible in a national assessment such as NAEP. The balance of multiple-choice to open-ended items was adjusted during item development when it was determined that important dimensions of the framework could not be captured appropriately with multiple-choice item types.

[35] Educational Testing Service, handout on distribution of assessment time by item format in 1994 and 1992, NAEP Network meeting, June 18, 1993.

Table 3.11. ***The 1992 NAEP reading assessment: Percentage distribution of assessment time by item format***

	Open ended	Multiple choice
Required by specifications	40	60
Actual grade 4	61	39
Actual grade 8	73	27
Actual grade 12	74	26

SOURCE: Educational Testing Service, handout on distribution of assessment time by item format in 1994 and 1992, NAEP Network meeting, June 18, 1993.

Assessment time. As described above, the Reading Framework specified longer item blocks so that students would have more time to read each passage and respond to its associated items. This was operationalized in the 1992 reading assessment by extending the block length to 25 minutes as compared to the 15-minute block length used in previous reading assessments. In addition, there was a small number of 50-minute blocks fielded at grades 8 and 12; these included the NAEP reader blocks in which students were allowed to select the passage they wished to read.

Item difficulty. An item-mapping study, parallel to the one described previously for mathematics items, was carried out for reading.[36] Like the mathematics study, the reading item-mapping study revealed deficiencies in the upper ranges of the scale where advanced achievement must be measured.

In summary, the Panel concludes that the 1992 reading assessment incorporated significant positive advances including a framework consistent with current theory and practice, interesting and authentic passages, longer testing time per passage, and a high proportion of open-ended items. However, more work appears to be needed to develop personal-response and critical-stance items broadly acceptable to the reading community. In addition, items that tap the most advanced levels of achievement are inadequately represented in the assessment.

Directing the NAEP Development Effort

Although they have been discussed separately, framework development and item construction are actually part of a single process that must flow seamlessly if the

[36] D. Pearson and L. DeStefano, "An Evaluation of the 1992 NAEP Reading Achievement Levels, Report Three: Comparison of Cutpoints for the 1992 NAEP Reading Achievement Levels with Those Set by Alternate Means," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993).

quality of the NAEP assessment is to be upheld. In addition, reporting, particularly if it is to include substantive achievement levels, must also be closely coupled with other aspects of the assessment. In the opinion of the Panel, the recent rapid expansion of the NAEP program has strained the project's resources, led to unfortunate discontinuities, and left insufficient time for careful deliberation. Not only have the costs and logistical complexities of a trial state assessment taken their toll, but the rapid development of new frameworks, plus the ambitious effort to launch achievement levels, all have exacted a price in terms of consensus building and thoughtful innovation. Participants in the framework-development processes for both mathematics and reading listed the rushed project schedules as their greatest complaint. And, given that different (although sometimes overlapping) groups of content experts were involved at different stages of the process that produced the assessments—framework development, item development, item review, and the development of descriptors to be used in reporting—professionals in both fields expressed profound concerns about the absence of a consistent stewardship to oversee the implementation of the framework vision. Similar problems with continuity are seen to exist across assessments, with no body of content experts having oversight over the evolution of the assessment.

These concerns are essentially the same as those expressed by the Panel in its evaluation of the content validity of the 1990 TSA. Some improvement was noted in that there was about one-third overlap between the Reading Framework planning committee convened by CCSSO under contract to NAGB and the 1992 reading instrument-development committee commissioned by ETS. The latter individuals meet four or five times over a 2-year period to consider the items at different stages of their development and to assist in the development of the scoring guidelines for the open-ended questions. A small number of these latter individuals are also being continued as members of the instrument-development committee that is currently overseeing the reading item-development and scoring process for 1994. While this overlap is better than none, it still falls far short of the recommended model, in which standing subject-matter panels, with long term responsibility for all aspects of the assessment-development process, would be established for each subject area.

Summary and Recommendations

1. The framework for mathematics, one of the two subjects included in the 1992 TSA, was developed in 1988 and used for both the 1990 and 1992 assessments. The Panel reviewed this framework in conjunction with its evaluation of the 1990 TSA and concluded that it represented a reasonable compromise between current instructional practices and the standards being put forward by the professional mathematics community at the time. Although the framework was held constant, the mathematics item pool was adjusted for the 1992 assessment to bring it into better alignment with the NCTM Standards (published in final form in 1989). In the view of the Panel, these changes were appropriate, given the rapid evolution of mathematics education during the intervening time period.

2. Compared to 1990, the 1992 mathematics item pool contained substantially more items overall (e.g., 45 percent more at grade 4). The numbers of difficult items and open-ended items were also increased, and two new item types were added—extended-response items and items that used cardboard geometric shapes, or manipulatives, to help the student reason through the problem presented. All of the changes were in the direction recommended by the mathematics experts who evaluated the content validity of the 1990 TSA on behalf of the Panel.

3. When items from the grade-4 mathematics assessment were independently classified by mathematics experts commissioned by the Panel, there was a high rate of agreement with the official NAEP *content* classifications (89 percent). Agreement was somewhat lower (69 percent) for the classification by *process*. In general, the disagreements highlight the shortcomings of the matrix design used for the 1992 Mathematics Framework, which required that each item be associated with one and only one cell of the framework.

4. In the judgment of the Panel's mathematics experts, nearly 60 percent of the 1992 fourth-grade items were related in some nontrivial way to one or more of the NCTM Standards themes. This represents an improvement over the 1990 grade-8 item pool, in which about 50 percent of the items had been judged by a similar panel to match any of the major themes of the standards. Only 9 percent of the 1992 grade-4 items, however, were judged to be *exemplars* of the standards themes. The mathematics experts generally reacted favorably to the new extended-response and manipulative items but found that the majority of the items designed to be solved with the aid of a calculator involved only uninteresting and unnecessarily tedious calculations.

5. The framework for reading, the second subject included in the 1992 TSA, was developed in 1990. This framework was well grounded in current reading research and included several innovative features not present in earlier NAEP Reading Frameworks, such as a definition of a "good reader" based upon a constructivist view of learning, use of authentic texts, assessment of reading across different reading situations, attention to the various cognitive aspects of reading, a high proportion of open-ended assessment items, and more time for students to respond to each reading passage. Reading experts surveyed in fall 1993, shortly after the release of the 1992 reading results, were reasonably satisfied with the framework, and most gave it a grade of "B" or higher.

6. Based upon the recommendations of the consensus project that developed the mathematics objectives, the reading consensus project added some useful enhancements including a set of detailed test specifications that elaborated upon the less technical framework document and assured better continuity between the framework and

the resultant item pool. Like the mathematics project, however, the reading project suffered from severe time constraints that allowed little time for meaningful discourse and consensus building in the wider professional community and precluded the project from undertaking any significant advances in measurement methodology.

7. Reading experts commissioned by the Panel reviewed the 1992 reading items for all three grades and independently classified the items according to the reading situations and aspects of reading specified by the framework. These experts judged the item distribution across ***reading situations*** to be a reasonable approximation of the ideal described in the framework, but concluded that the assessment was significantly lacking in items that adequately measured the *critical stance* or *personal response* aspects of reading.

8. A very high percentage of the 1992 reading items utilized an open-ended format, with open-ended items reaching 74 percent of assessment time at grade 12. In addition, item blocks were extended from 15 minutes to 25 minutes so that students would have more time to read each passage and respond to its associated items. These innovations reflect NAEP's efforts to align with current concepts in the subject-area professional communities.

9. A separate item-mapping study undertaken by the Panel revealed insufficiencies in the upper difficulty range of the 1992 mathematics item pool despite the fact that the NAEP contractor had attempted to rectify this problem between 1990 and 1992. Similar deficiencies were observed in the 1992 reading item pool. The Panel therefore concludes that the 1992 assessment, although well designed and innovative in many respects, was still inadequate for measuring very advanced achievement on challenging subject matter.

10. There was about one third overlap between the planning committee that wrote the 1992 Reading Framework and the instrument-development committee that reviewed the reading assessment items at various stages of their development and assisted in the creation of scoring guides for the open-ended questions. A small number of these latter individuals are also being continued as members of the instrument-development committee for the 1994 reading assessment. While this overlap is better than none, it still falls far short of the model recommended by the Panel in its first and second reports, in which standing subject-matter panels with long term responsibility for all aspects of the assessment-development process would be established for each subject area.

In light of these findings, the Panel makes the following recommendations for the assessment-development process in general and the content of the mathematics and reading assessments in particular.

The Panel begins by repeating its recommendation from the previous evaluation of the TSA that a structure be established to permit longer and more thoughtful planning cycles for the assessment:

- *planning must begin well in advance for future assessments so that adequate time can be allowed to develop exemplary assessment tasks and to find ways to measure constructs not well covered by today's assessment methodologies. ... One-year-at-a-time reauthorizations and extensions just prior to the scheduled field tests are not adequate. Congress should authorize NAEP assessments several years prior to their scheduled administration in order to resolve this ongoing problem [of overly tight timelines].*[37]

In addition, as discussed in chapter 10, standing subject-matter panels should be established to provide continuity within and across assessment cycles.

With regard to item content, the Panel recommends the following:

- *Increase efforts to include more difficult and challenging items in all subject areas. For efficiency, a modified design in which these items would be administered to selected samples of high-performing students could be considered.*

- *Continue the trend toward the extensive use of open-ended and other performance-type items. However, research should be undertaken to better understand the attributes of these item types and their impact on the measurement of student performance.*

- *Investigate parameters of the test situation that can be manipulated to encourage more thoughtful responses to the extended-response items. For example, three to five extended-response items might be organized into a single block with a 25- to 30-minute time limit.*

- *Utilize special studies to address skills not readily captured with short response times or paper-and-pencil assessments. For example, in mathematics, NAEP might use trained observers to evaluate a student's use of manipulatives to solve problems or, in reading, NAEP could provide time in advance of the assessment for a student to read and reflect upon materials included in the assessment.*

For mathematics,

- *Increase the numbers of items measuring themes from the NCTM Standards, particularly the themes of communications and connections.*

[37] The National Academy of Education, *Assessing Student Achievement*, 29-30.

- *Replace the current content-by-process matrix, which implies that each item can be uniquely classified into a single content-by-process category, with a model that better represents the integrated nature of mathematical thinking. For example, items that make connections across mathematical content areas such as geometry and algebra should be encouraged, and individual items should tap a range of mathematical processes such as problem solving and communicating about mathematics.*[38]

- *Allow unrestricted use of calculators on all but trend blocks or specifically excluded items.*

- *Modify both the time limits and wording of items involving manipulatives to give students more opportunity to explore alternative solutions. A partial-credit model should be used in scoring these items.*

For reading,

- *Increase efforts to design critical-stance and personal-response items that will be widely accepted by the professional reading community.*

[38] The 1994 Mathematics Framework appears to have been modified significantly in this regard. For example, the matrix framework has been discontinued in favor of a model consisting primarily of the five content strands used in the 1990 and 1992 assessments, and items can be classified into more than one content strand if they address substantive content from more than one area.

4 *Reporting and Impact of State NAEP Results*

Ultimately, the purpose of NAEP is not only to collect data on student performance, but also to make these data available in a manner that is accurate, accessible, and useful. With the advent of the TSA and the push toward standards-based reporting, both the magnitude and the complexity of this reporting task have expanded greatly. One impact of the TSA has been the sheer volume of data to be reported. Overall performance, performance by demographic subgroups and subject-matter subscales, and the relationship of performance to teaching practices and student habits, all must now be provided for individual states—singly and in comparison to one another—as well as for the nation as a whole.

Beyond this, there has been a profound secondary impact of the TSA on reporting. The major argument for reporting NAEP results at the state level was that education change could best be effected by providing states with information on their own educational progress—over time and, to a lesser extent, in relation to other states.[1] For this premise to hold, however, the information to states must be provided in a useful manner. Hence, the advent of state NAEP has led to much greater concern with the scope and quality of NAEP reporting. For example, individualized reports have been prepared for each state so that findings will not have to be culled from cumbersome compendia. Significant attention has also been paid to the appropriate reporting of state comparisons, as well as to the preparation of special focused reports that probe the implications of NAEP results for various education choices.

In 1990, the TSA involved one grade and one subject, and no trend data were available. For 1992, there were two grades and two subjects to report, and, in the case of eighth-grade mathematics, the reporting also had to accommodate comparisons between 1990 and 1992. In addition, there was strong pressure to release each set of results as soon as they were available, sometimes in response to an externally imposed deadline such as the publication schedule of the National Education Goals Panel (NEGP) Report Card. At the same time, unresolved technical issues, particularly those surrounding the shift to standards-based reporting, worked to slow down the pace of the reporting. The reporting schedule for the subjects included in the 1992 TSA, as determined by the date of this report, is shown in figure 4.1. In addition, NAEP will be releasing reports on various data-collection efforts that occurred only at the national level in 1992. These include the results of a full national writing assessment at three grades and an updated trend report covering reading, writing, mathematics, and science.

[1] By contrast, it has generally been felt that national NAEP statistics have only limited potential for direct policy impact because few of the relevant education decisions are made at the federal level.

Figure 4.1. ***Data-release schedule for subjects included in the 1992 NAEP TSA***

Release date	Subject	Results
January 12, 1993	Mathematics	Highlights of national results—grades 4, 8, and 12
April 8, 1993	Mathematics	Full national and state results—state results at grades 4 and 8 only*
April 8, 1993	—	*Interpreting NAEP Scales,* a report on the evolution of NAEP reporting with particular emphasis on the new achievement levels
September 1, 1993	Mathematics	*Can Students Do Mathematical Problem Solving?,* a focused report on constructed-response items—national and state results
September 15, 1993	Reading	Full national and state results—state results at grade 4 only*
To be announced	Mathematics	Focused report on effective schools and instruction
To be announced	Reading	Focused report on a special fourth-grade reading study involving oral reading and response data and portfolio-like samples of classroom reading activities (national sample only)
To be announced	Reading	Focused report on constructed-response items and performance by type of reading situation
To be announced	Mathematics	Focused report on the relationship of the NAEP results to the NCTM Standards

* Full national and state results include a *Report Card* and a *Data Compendium* for the nation and the states, an *Executive Summary* of the *Report Card,* and individual *State Reports* and *Data Almanacs* for each participating state. *Data Compendia* have not been available until shortly after the main release date.

Reporting the 1992 TSA in Mathematics

Although 1992 marked the second TSA in mathematics, many aspects of the reporting structure had to be newly designed to accommodate the trend comparisons between 1990 and 1992 and the fact that there were now two grades to report rather than just one. The 1990 reports, although generally well received, had already drawn criticism for being too densely filled with information; yet the 1992 reports had to accommodate two to three times as much data. A typical 1992 state report was 212 pages, some 70 pages longer than the report for the same state in 1990. However, the *NAEP 1992 Mathematics Report Card for the Nation and the States* was only 368 pages,[2] some 160 pages less than its 1990 counterpart, *The STATE of Mathematics Achievement*.[3] The savings were attained by a better integration of national and state results and by moving much of the information from the NAEP student and teacher questionnaires to a separate focused report (see figure 4.1). In addition, a new document, called a *Data Compendium*, was produced to accommodate some detailed information that was of less general interest.

The first TSA had demonstrated the problems inherent in making state-to-state comparisons that were both statistically valid and readily interpreted by a lay audience. Many news articles published state rankings without specifying the statistical significance (or nonsignificance) of the differences portrayed. Moreover, journalists rejected as too complex a graphic that portrayed significant state-by-state differences using a format similar to the height-by-weight charts that guide pantyhose consumers to the correct choice of size.

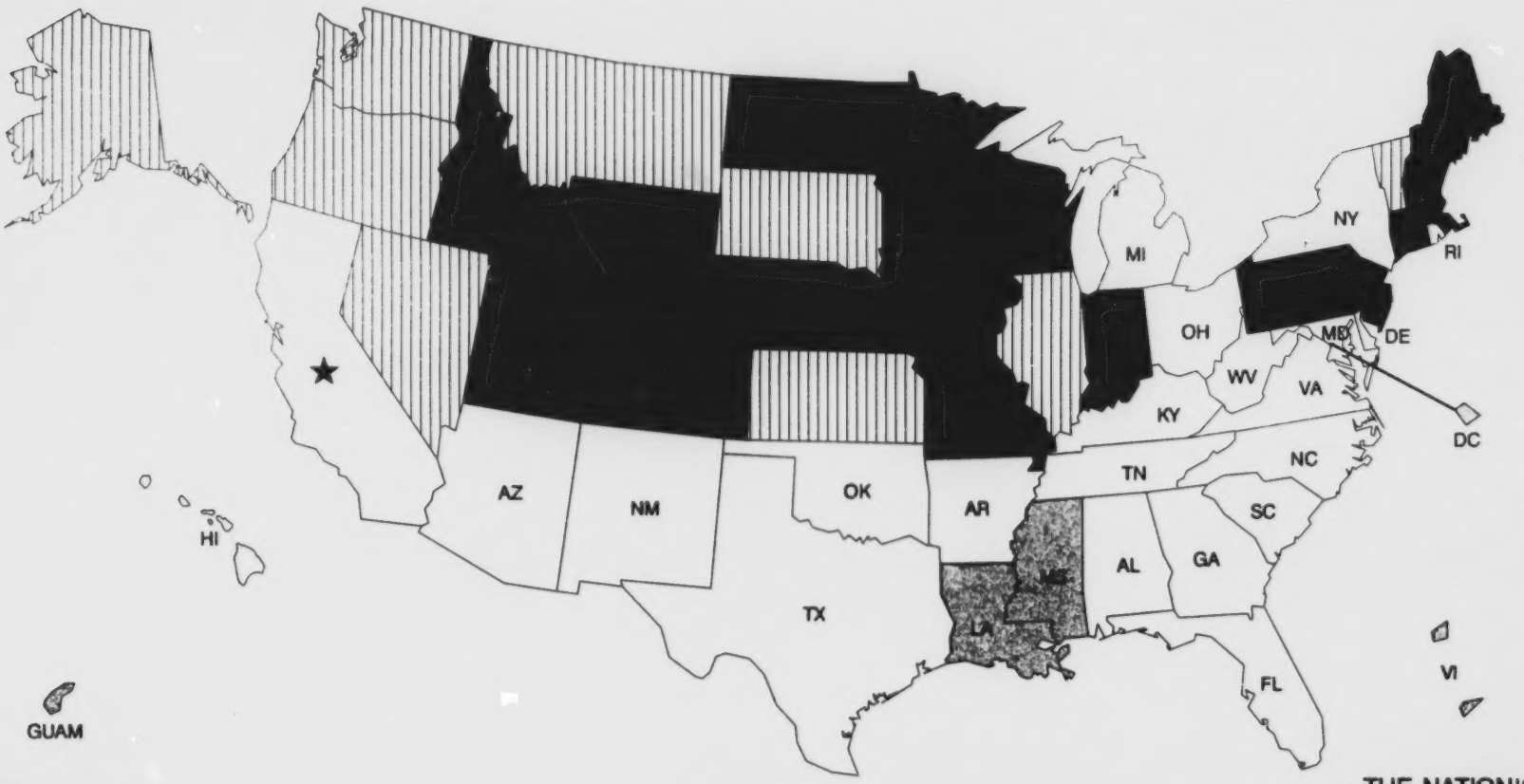
Consequently, for 1992, several new devices were introduced to further assist in the portrayal of state comparisons. First, each state report contained a customized U.S. map that used different fill-patterns to clearly show which states performed significantly better than, the same as, or worse than the target state (figure 4.2). (A similar map was used in the *Report Card* to indicate which states had improved significantly since 1990.[4]) A second new reporting device was the use of quintile rankings to concisely summarize comparative state performance on several dimensions including gender, race-ethnicity, and the five mathematics content areas. Finally, a new series of prose paragraphs were included in the press kits that customized, for each state, specifically what should be said to accurately portray that

[2] I.V.S. Mullis, J.A. Dossey, E.H. Owen, and G.W. Phillips, *NAEP 1992 Mathematics Report Card for the Nation and the States* (Washington, D.C.: National Center for Education Statistics, April 1993).

[3] I.V.S. Mullis, J.A. Dossey, E.H. Owen, and G.W. Phillips, *The STATE of Mathematics Achievement* (Washington, D.C.: National Center for Education Statistics, June 1991).

[4] Among the 37 states and territories that participated in both 1990 and 1992, 18 showed statistically significant increases in performance and 19 remained the same. No states or territories showed statistically significant decreases in performance.

Figure 4.2. ***Comparisons of overall mathematics proficiency at grade 8: California***



SOURCE: National Center for Education Statistics, *NAEP 1992 Mathematics State Report for California* (Washington, D.C.: Author, April 1993), 17.

state's ranking in relation to all other states. These innovations were in the spirit of the Panel's recommendation, in its first report, that "NCES continue to provide the media with specific examples of interpretations that are (or are not) supported by the results as well as continue efforts to construct new tables and graphs that effectively display the data."[5]

The Panel also noted that, for the first TSA, "[press] coverage of some of the most interesting and potentially useful aspects of [NAEP] contextual data, such as the relationship of student performance to differences in mathematics curriculum or teacher preparation, was scant at best," and suggested that "NCES consider spreading the release of the data over several months."[6] This suggestion is consonant with the NCES decision to generate a series of focused reports for 1992 and issue them on a staggered schedule following the main release of the overall results in each subject area. The focused reports appear to be a useful innovation in that they provide more amplitude for describing the educational and policy implications of the data; it is somewhat problematic, however, that the release of these reports—originally scheduled to occur at monthly intervals—has been delayed.

Reporting by Achievement Levels

The reporting of the 1992 assessment marked the first time that the NAEP achievement levels were used as the primary reporting device. The achievement levels define within-grade-level cutscores on the NAEP scales at which student performance meets the criteria for basic, proficient, or advanced achievement. Prose descriptions of behaviors appropriate to each achievement level also are included; these descriptions are primarily intended to articulate appropriate content goals for student attainment.

Considerable controversy attached to the decision to report by achievement levels, and the reporting format went through a number of permutations before it was finalized. NAGB had directed that the achievement levels be used as the primary basis for reporting, but this Panel, in its February 1992 report on the 1990 TSA, concluded that the utility of the achievement levels had not yet been confirmed and strongly suggested that "NCES report the results from the 1992 assessment using the NAEP scale score and its anchor levels, whether or not the 1992 achievement levels [were] used."[7] In May 1992, the assessment subcommittee of CCSSO's Education Information Advisory Committee (EIAC), which represents one of state NAEP's key stakeholder groups, the state assessment directors, weighed in with its position. It passed resolutions calling for further validation of the achievement levels but also recommended that the results of the 1992 TSA in mathematics be presented primarily in terms of achievement levels, with the traditional anchor-level data also presented in the report, perhaps in an appendix. NCES prepared its reports in accordance with the

[5] The National Academy of Education, *Assessing Student Achievement in the States* (Stanford, CA: Author, 1992), 52.

[6] Ibid., 53.

[7] Ibid., 39.

EIAC recommendations, citing as its rationale for the inclusion of the anchor levels the precedent of other statistical agencies that have continued to report both old and new measures during times of transition.

Further difficulties arose during the final adjudication of the 1992 reports, when these reports, like all NCES reports, went through a process of internal review by U.S. Department of Education staff. Once these reviewers became aware that the behaviors described in the achievement levels were not closely aligned with students' actual test performance, they were unwilling to sign off on the reports.[8] At the point that these concerns arose, however, highlights of the national mathematics results had already been released using the achievement levels. After further consideration, therefore, and in the interest of hastening the release of the already-delayed state results, the full state and national results were eventually released as prepared. Disclaimers were included, however, that explained the limitations on the inferences that could be drawn from the current achievement levels as well as the agency's reasons for choosing to report them, namely that "The public needs the data in this report to see for themselves what standards-based reporting might do and to evaluate the often conflicting claims of adherents and detractors... ."[9] In addition, the report *Interpreting NAEP Scales* was released concurrently with the full national and state results to help consumers understand the strengths and limitations of each of the methods for reporting.[10]

State Response to Achievement Levels and Other New Report Features

Preliminary results from a survey of state assessment directors and mathematics specialists conducted during the summer of 1993 indicated continuing support for the use of achievement levels. *More than half of the respondents from states that participated in the 1992 TSA considered the achievement levels to be, overall, more useful than the older anchor levels, while only 7 percent considered them less useful.* Further, among those who had a clear opinion, the achievement levels were nearly three times as likely to be rated "easier to understand" and more than five times as likely to be rated "more relevant for communicating the state of education" and "more likely to have an impact on education policy in the respondent's state." Perhaps because they were already familiar with the issues, only about a third of the respondents reported reading the *Interpreting the NAEP Scales* report in any level of detail. Those that read it considered it useful for their own understanding, but virtually no one passed it on to their constituents or used it in reporting to the press.

[8] For a more complete discussion of this problem, see The National Academy of Education, *Setting Performance Standards for Student Achievement* (Stanford, CA: Author, 1993).

[9] Mullis et al., *NAEP 1992 Mathematics Report Card*, 34.

[10] G.W. Phillips et al., *Interpreting NAEP Scales* (Washington, D.C.: National Center for Education Statistics, April 1993).

The state respondents were very favorably impressed by the maps portraying comparisons between states and over time. Among respondents from participating states, about two thirds evaluated these new graphics as "very useful," particularly for helping to present the results to constituent groups within the state. The new quintile tables, focusing on comparisons among subgroups, were necessarily less central to the main thrust of the reporting. Nevertheless, they generally received "somewhat useful" ratings from the state respondents.

The overall impression of the reports was favorable, with nearly all respondents evaluating them as good as or better than the well-received 1990 reports. Among those respondents whose states had participated in both 1990 and 1992, about 40 percent found the 1992 reports to be more useful, primarily because of the innovations discussed above.

Reporting the 1992 TSA in Reading

The reading reports parallelled the mathematics reports in most respects, and all of the reporting innovations designed to facilitate appropriate state-by-state comparisons were retained. The primary difference was in the treatment of the achievement levels. Based on a compromise that NCES and NAGB reached during the adjudication of the mathematics reports, the behavioral-anchoring technology, previously used to give meaning to the traditional NAEP anchor levels, was applied to the new achievement levels for reading. Using data from items answered correctly by at least 65 percent of students at a given achievement level, reading experts prepared prose descriptions of what students scoring at each achievement level actually could do. The descriptions were incorporated into the main body of the report, where they were contrasted with the official achievement-level descriptors emphasizing aspirations for what a student *ought* to accomplish in order to be judged basic, proficient, or advanced. An example for grade-4 basic is given in figure 4.3.

Figure 4.3. ***Fourth-grade basic achievement level in reading***

Percentage at or above basic: 59 (1.1) scale cutpoint: 212	
Achievement-level description	Anchoring description
Fourth-grade students at the Basic Level should demonstrate an understanding of the overall meaning of what they read. When reading text appropriate for fourth graders, they should be able to make relatively obvious connections between the text and their own experiences.	Fourth-grade students at the Basic Level could understand uncomplicated narratives and high-interest informative texts. They identified obvious themes, located explicit information, summarized parts of the text, and made judgments about character's actions.

SOURCE: I.V.S. Mullis, J.R. Campbell, and A.E. Farstrup, *NAEP 1992 Reading Report Card for the Nation and the States* (Washington, D.C.: National Center for Education Statistics, September 1993), 27.

The Panel considers the 1992 reading reports, like the mathematics reports, to be of generally high quality, although the Panel does not support the use of the current achievement levels on technical grounds.[11] *Given that the achievement levels were used, the use of the anchoring descriptions to add meaning to the reported cutscores was appropriate.*

Delays in the Reporting Schedule

Despite generally favorable impressions of the reports per se, one of the major perceived problems with NAEP, and particularly with the state assessment component, concerns the seemingly unavoidable, lengthy delays between the assessment administration and the reporting of results. Time diminishes both the impact and the policy relevance of reported results, and state NAEP in particular struggles under a paradox: compared to national NAEP, it is aimed more directly at those policymakers who desire timely data, and yet, due to the greatly increased data-analysis burden and the tremendous complexities of such undertakings as customized, computer-generated state reports, more time is required to make state data available. For the 1990 TSA, reports were not released until June 1991, some 15 months after the assessment was administered. NCES had hoped to considerably improve this turn-around time for the 1992 TSA and to make at least the mathematics portion of the results available some time around December 1992. Largely due to issues and concerns surrounding the achievement levels, however, this did not prove feasible, and the reporting schedule for the 1992 mathematics results was gradually delayed to April 1993. In the meantime, highlights of the national results were presented to the press in a brief data summary in mid-January. Finally, the three focused reports in mathematics, originally scheduled to be released at monthly intervals following the main reports, were further delayed as attention shifted to reports of the reading results. The first focused mathematics report was issued at the beginning of September 1993, and the other two are still pending as of this writing.

With regard to reading, it was originally anticipated that the 1992 results would be released around June 1993, placing them no later (in relation to the assessment administration) than the 1990 mathematics results. However, the delays in the 1992 mathematics reporting schedule, along with the refinements to the achievement-level reporting process described above, subverted this plan. At one point it seemed as though the reading results might be delayed by as much as 5 months. NAGB considered this unacceptable because the reading data were slated to be included in the third annual NEGP Report Card, to be issued at the end of September. Accordingly, and by virtue of much hard work on the part of NCES and the NAEP contractor, a set of reading reports—nearly as complete as the April mathematics reports—were readied for release on September 15, 1993. As of this writing, the main reading reports are scheduled to be followed by focused reports on constructed-response items and types of reading situations and on the special reading study at grade 4. The first of these, however, is not scheduled to be released until at least December 1993.

In general, the Panel finds that the reporting of the 1992 state assessment results was delayed, both overall and in terms of the time elapsed between related reports.

[11] See chapter 8 in this report and also The National Academy of Education, *Setting Performance Standards.*

1992 Dissemination Activities

As it did for the 1990 results, NCES held national press conferences in April and September 1993 to present the 1992 mathematics and reading reports. The level of coverage provided by NCES was quite similar for the 1990 and 1992 TSAs, with the Secretary of Education, NAGB board members, and representatives from the mathematics and reading education communities providing commentary on the results, and a live satellite feed of the press conference available to interested states.

NCES introduced a new format for disseminating the results of the focused reports. On October 19, 1993, about 6 weeks after the release of the first focused report *(Can Students Do Mathematical Problem Solving? Results from Constructed Response Questions in NAEP's 1992 Mathematics Assessment)*,[12] NCES hosted a half-day seminar at which panel members representing curriculum development, teaching, educational research, and educational measurement presented their perspectives on the results. The conference was well attended, with representatives from the Mathematics and Science Education Board, NCTM, the National Science Foundation (NSF), and the American Association for the Advancement of Science, as well as about a dozen teachers, assessment coordinators, and curriculum supervisors from the surrounding school districts.

Preliminary analyses of the Panel's survey data suggest that states were somewhat less active in disseminating the TSA results for 1992 than they were for 1990. When the 1990 results were released, 97 percent of participating states issued a press release; 65 percent held a press conference. For the 1992 mathematics results, press releases were again widespread (95 percent of participating states), but only 29 percent held a press conference. A number of respondents attributed this diminished effort to idiosyncratic reasons such as staff or office restructuring or competing priorities, but others suggested that there was simply less interest now that the TSA was not a novelty. The Panel is continuing to analyze the data to determine, for example, whether dissemination was related to such factors as the state's performance on the TSA, state education reform activities, or the perceived value of the TSA to the state. The Panel also will continue to collect and analyze data on more specialized dissemination efforts carried out by the states, such as the use of NAEP results in position papers, professional presentations, or inservice programs.

Impact of the TSA

Overall, indications suggest that state NAEP continues to sustain or even to increase its perceived value to its main constituents, state-level educators and policymakers. *Perhaps the most compelling evidence for this is the fact that sign-ups for the state assessment program increased from 1990 to 1992 and appear likely to increase again*

[12] J.A. Dossey, I.V.S. Mullis, and C.O. Jones, *Can Students Do Mathematical Problem Solving? Results from Constructed Response Questions in NAEP's 1992 Mathematics Assessment* (Washington, D.C.: National Center for Education Statistics, August 1993).

for 1994. The latter is all the more remarkable in the face of the lingering uncertainties about funding levels, and consequently about the grades and subjects to be assessed, that plagued the program well into the time when school recruitment for 1994 should have begun.

In addition, early analyses of the 1993 survey data, collected 3 to 4 months after the main release of the mathematics results and before the release of the reading results, indicate that nearly half of the state assessment directors and mathematics specialists from participating states found the TSA to be a "very worthwhile" return for the effort; almost all of the remainder judged it to be "somewhat worthwhile." Also, as a consequence of the trend data—available for the first time in 1992 (albeit for a very short, 2-year trend), *nearly half of the respondents indicated that they found the TSA more valuable in 1992 than in 1990.*

Where exactly does this value reside? Proponents of the NAEP state assessments as an agent of change have focused both on the public, whom these proponents hope to see aroused to a more knowledgeable and concerned stewardship of their children's education, and on educators and policymakers, each of whom, it is hoped, will hold themselves accountable in the face of this tangible evidence of the quality of their state's educational outcomes. At the same time, the NAEP program increases in value to the extent that educators can glean from its findings, not just overall measures of their success but also evidence suggestive of the *types* of reforms that can improve educational performance.

Press Coverage

Press coverage is one way of disseminating the data and increasing the program's impact, particularly on the public, and consequently on policymakers. However, as one might expect, extensive press coverage is likely to be highly contingent upon novelty and "newsworthiness" and therefore is difficult to sustain over repeated cycles of assessment. This would seem to be borne out by the coverage of the 1992 mathematics results. *Compared to 1990, there were fewer stories in the top 50 newspapers nationally, and those stories that did appear were not as lengthy or detailed.* To some extent this may have been due to the fact that the national results had already been previewed in January, or to the fact that the states themselves put only limited energy into promoting their story to the press. But it is also true that the "wake-up call" for American education had already been sounded 2 years earlier and might not ring quite so loudly a second time. Coverage of the NAEP reading and mathematics reports released in September was also somewhat minimal, which would tend to support the latter hypothesis.[13]

Widespread dissemination via the press may be difficult to sustain through repeated, frequent cycles of testing. Evidence from the surveys and case studies carried out by the

[13] As further evidence for the novelty hypothesis, the Panel notes that the NCES adult literacy study, also released in September, received widespread coverage for its findings that substantial proportions of adult Americans were handicapped by inadequate reading and mathematics skills. This was the first adult literacy study to be reported since 1975.

Panel after the 1990 TSA and discussed below, however, suggest that impact can be achieved by focusing dissemination efforts on those policymakers and practitioners most capable of effecting change.

Evidence of Impact from the Reporting of the 1990 TSA

The Panel is continuing its ongoing investigation of the impact of the reporting of the TSA. At the time of this writing, a survey of the short term impact of the 1992 mathematics reports has just concluded, and additional surveys and case studies for the short term impact of the reading reports and the longer term impact of both subjects are planned. In the meantime, the best evidence concerning the ability of state NAEP to effect education change comes from the surveys and case studies carried out during 1992 to assess the longer term impact of the 1990 TSA.[14]

Results of these analyses suggest that the 1990 TSA carried weight, not in isolation, but because it articulated well with other contemporary influences in mathematics education, particularly the NCTM Standards. At the time that the 1990 results were released, nearly all of the states were involved in some level of curricular or assessment reform affecting mathematics. The Panel's January-February 1992 survey showed that, among the states that were making such changes *and* had participated in the 1990 TSA, 46 percent of respondents indicated that the changes were being influenced by NAEP and the TSA (table 4.1). Perhaps surprisingly, even in nonparticipating states, 30 percent of the respondents indicated such an influence. Both proportions were up from the Panel's previous survey, in July-August 1991, when influence was reported by only 33 percent of respondents from participating states and 13 percent of respondents from nonparticipating states.

[14] F.B. Stancavage, E.D. Roeber, and G.W. Bohrnstedt, "Impact of the 1990 NAEP Trial State Assessment Program: A Followup Study," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

Table 4.1. ***Percentages of respondents reporting 1990 TSA influence on decisions to change mathematics curriculum, instruction, teacher preparation and/or influence on what specifically is being changed****

	Respondents from participating states	Respondents from nonparticipating states	Total
July-August 1991 (P = 58, NP = 23)	33	13	27
January-February 1992 (P = 65, NP = 23)	46	30	42

Note: P=participating, NP=nonparticipating

* Survey included all states. Table includes states making changes only. In almost all cases there were two respondents per state—the state assessment director and the state mathematics specialist.

SOURCE: F.B. Stancavage, E.D. Roeber, and G.W. Bohrnstedt, "Impact of the 1990 NAEP Trial State Assessment Program: A Followup Study," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

Respondents credited NAEP and the 1990 TSA with adding impetus to already planned or desired changes, accelerating discussions of state goals, and increasing pressure to align with national standards. Some very specific uses and influences attributed to the 1990 TSA included tipping the balance in favor of using calculators in the classroom and using the sample NAEP items in teacher inservice sessions to illustrate the kinds of skills and activities teachers must learn to incorporate in their lesson plans.

As shown in table 4.2, a very similar picture is evident for the reported influence of the 1990 TSA on *state assessments*. Here the reported influences included hastening the passage of accountability legislation and the adoption of new or revised state assessments, the use of calculators on assessments, and the influence of NAEP items on assessment design.

Table 4.2. ***Percentages of respondents reporting 1990 TSA influence on decisions to change mathematics assessment program and/or influence on what specifically is being changed****

	Respondents from participating states	Respondents from nonparticipating states	Total
July-August 1991 (P = 54, NP = 19)	39	21	34
January-February 1992 (P = 57, NP = 15)	51	21	44

Note: P=participating, NP=nonparticipating

* Survey included all states. Table includes states making changes only. In almost all cases there were two respondents per state—the state assessment director and the state mathematics specialist.

SOURCE: F.B. Stancavage, E.D. Roeber, and G.W. Bohrnstedt, "Impact of the 1990 NAEP Trial State Assessment Program: A Followup Study," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

The Panel also carried out a series of nine case studies in the fall of 1992 in order to look in greater detail at the ways in which impact was affected by each state's unique needs and circumstances.[15] A few themes emerged strongly from these case studies, corroborating impressions that had already been formed based on the earlier structured interviews with national, state, and local respondents. The first of these themes is that the state departments of education are the primary consumers of state NAEP data. These departments mediate the ways in which the data are covered by the local press, use the data to inform their own planning efforts, and are primarily responsible for bringing the information to the attention of state policymakers and local educators. A second theme is that, for these state-level consumers, NAEP is primarily valued for its role in sustaining and supporting broader trends in education reform, particularly the movement toward institutionalization of the NCTM Standards, and, more broadly, the general movement toward accountability and standards in education. A third theme was that, not surprisingly, the greatest initial public impact occurred in states that performed worst on the 1990 TSA; however, even in these states, the enduring impact of the results was largely dependent on the actions of key personnel in the state departments of education. Of course, the actions of these individuals were in turn influenced by the availability of departmental resources, as well as by the extent to which competing efforts demanded their time and attention.

All of the case-study states were engaged in a process of ongoing education reform at the time that the 1990 TSA data were released. Several had received Eisenhower grants and/or NSF State Systemic Initiative awards to support innovation in mathematics and science, and all had made at least some efforts to bring curriculum,

[15] The nine case-study states were California, Colorado, Connecticut, Delaware, Florida, Indiana, Louisiana, Nebraska, and North Carolina.

instruction, and assessment in line with the recommendations of the NCTM Standards. The specific utility of the 1990 TSA data for a given state was at least partly dependent on how well these data articulated with the ongoing reform process in the state.

For example, in Florida, the impact of the 1990 TSA was dampened by the fact that the education community was deeply involved in the implementation of an education accountability law that had been passed in June 1990. Although this law mandated continued participation in NAEP as one dimension of accountability, the TSA mathematics results were somewhat tangential to the main thrust of the law, which was designed to foster local initiative and control by charging local school districts with assessing their needs and developing plans for meeting those needs. In addition, economic conditions had recently worsened in the state, exerting pressures on school budgets and diverting attention to financial issues.

By contrast, in Indiana, information from the 1990 TSA was valued because it dovetailed nicely with ongoing efforts to bring mathematics instruction and assessment into greater alignment with the NCTM Standards and to raise expectations for student performance. For example, prior to the release of the TSA data, Indiana's state assessment program had a validated cutscore, and students who fell below it were placed in a mandatory 80-hour summer remediation program. Six months after the release, the cutscore was revalidated, and the standards were raised. While the state superintendent had been promoting this change for some time, the TSA results influenced a state study leading to a new definition of "borderline" students and gave ammunition to the superintendent's recommendations.

Four of the nine case-study states, including the two strong local-control states in the sample (Colorado and Nebraska), credited the 1990 TSA with promoting greater use of calculators in the classroom. In Nebraska, for example, case-study respondents pointed out that shortly after the 1990 TSA was administered many more schools and classrooms began using both standard and graphing calculators, with some financial help from the state's NSF grant. At the same time, the University of Nebraska revamped preservice classes for elementary mathematics teaching to add calculators and also increased the number of preservice mathematics courses from one to three. Here as elsewhere, however, it is difficult to isolate a *specific* TSA influence, since many other factors were at work at the same time. Most notably, the Eisenhower grant was paying for a series of workshops in Nebraska's 19 Educational Service Units, designed to provide information on NCTM Standards and encourage teacher involvement in mathematics education reform. A large national conference, organized by NCTM in coordination with the Nebraska Association of Teachers of Mathematics, was also held in Nebraska in the spring of 1991.

Five of the case-study states used NAEP results and materials in inservice training programs put on by the state departments of education. One of the more interesting examples of such efforts was in Louisiana, where the Louisiana Public Broadcasting station approached the state mathematics supervisor and asked for her help in producing several hour-long segments on the teaching of mathematics and the state systemic initiatives program for a targeted audience that included the general public as well as students and teachers. These shows were a direct result of the publishing of Louisiana's 1990 TSA scores, which were among the lowest in the nation.

North Carolina is especially noteworthy among the case-study states (actually among all states) for the strong, proactive role it took in turning the NAEP materials and results to positive effect. North Carolina had been engaged in education reform prior to the 1990 TSA, and, based on other test scores, the impression was that the state was making acceptable progress. When the 1990 TSA results were released, however, their unhappy message sent shock waves. Contrary to expectations, the average proficiency score for the state's eighth graders was 250, lower than for students in any other state but one.

During July and August, immediately after the release of the TSA data, the assistant superintendent for research and development and the state assessment director met with every superintendent in the state in cluster meetings to discuss the TSA results, the new mathematics curriculum, and the directions in which mathematics instruction and assessment were headed. In January 1992, another series of meetings was held with superintendents, again focusing on the TSA results.

An action plan was developed, and one of its primary components was the development of a 1-hour video, highlighting the TSA results and their implications, which was broadcast in October and November via closed-circuit television to every school system in the state. The first half of the video, developed by the state assessment staff, used TSA data to substantiate the identification of problems. Six or seven areas were identified in which the state's students were doing poorly in comparison to the national sample (for example, in items involving the Pythagorean theorem or estimation). The TSA data from teacher questionnaires were used to describe classroom practices and relate classroom practices to student performance. The video developed the hypothesis that, although North Carolina teachers were following many of the practices commonly identified as supporting mathematics achievement (for example, using manipulatives), teachers might be focusing on arithmetic too much and setting expectations too low. It also appeared that students were too frequently placed in ability groups, and even those in the top groups were not doing as well as expected. The emphasis was on identifying problems, not placing blame. The second half of the video, developed by the state curriculum and instruction staff, suggested solutions to the problems identified by the first half. It included samples of what the state mathematics curriculum would be like if properly implemented.

In addition, the state mathematics director, noting that teachers needed additional training before they could reasonably be expected to implement the suggested changes, developed training for team leaders, who are in turn training others in school districts across the state—the "trainer-of-trainers" model. The ripple action of this training should eventually reach every mathematics teacher in North Carolina.

In summary, the lessons from the telephone surveys and case studies conducted to date suggest that the TSA has had a positive impact on curriculum and assessment in many of the states that chose to participate, as well as some that did not. Further, the extent of this impact is likely to grow so long as state NAEP maintains its credibility as an "outside authority" that can be used to bolster the messages of state policy, curriculum, and assessment leaders. Tradeoffs and linkages between NAEP and state assessments are still evolving, and clearly the course of this evolution will greatly affect the ways in which state NAEP data are used in the future. For now, it appears that the state-level impact of NAEP is inversely related to the investment that the state has in its own assessment system.

1. In summary, the Panel concludes that the reporting of the 1992 TSA was successful. State department of education employees, the primary consumers of TSA reports, indicated a generally favorable overall impression of the 1992 reports, judging them as good as or better than the well-received 1990 reports. Among those respondents whose states had participated in both 1990 and 1992, about 40 percent found the 1992 reports to be more useful, primarily because of a number of well-received innovations including easy-to-read maps that compared performance between the states or between 1990 and 1992.

2. More than half of the respondents from states that participated in the 1992 TSA considered the achievement levels to be, overall, more useful than the older anchor levels, while only 7 percent considered them less useful. Further, among those who had a clear opinion, the achievement levels were nearly three times as likely to be rated "easier to understand" and more than five times as likely to be rated "more relevant for communicating the state of education" and "more likely to have an impact on education policy in the respondent's state." The Panel endorses standards-based reporting but opposes the use of the current achievement levels on technical grounds. Given that the achievement levels were used, the Panel endorses the use of the anchoring descriptions to add meaning to the reported cutscores.

3. Despite the high quality of the 1992 reports, the reporting process has proved cumbersome and lengthy to implement. State mathematics results were not available until April 1993, fully 13 months after the assessment, and reading results were delayed another 5 months, until September 1993. Moreover, the release of the focused reports, designed to elaborate upon specific findings of interest, such as the relationship of NAEP scores to teaching practices, has been delayed far beyond the schedule originally projected. The Panel therefore finds that the reporting of the 1992 state assessment results was unacceptably delayed, both overall and in terms of the time elapsed between related reports.

4. Fewer participating states held press conferences for the 1992 results, and press coverage of the 1992 TSA results was neither so extensive nor so detailed as the coverage given to the first TSA in 1990. Widespread dissemination via the press may be difficult to sustain through repeated, frequent cycles of testing. Evidence from the surveys and case studies carried out by the Panel after the 1990 TSA, however, suggest that impact has been achieved by focusing dissemination efforts on those policymakers and practitioners most capable of effecting change. For example, impact has resulted from using NAEP data in support of arguments to the state legislature or from building teacher inservice programs around released NAEP items and NAEP results.

5. The main constituencies for the TSA, state educators and policymakers, do appear to continue to value the program. Perhaps the most compelling evidence for this is the fact that sign-ups for the state assessment program increased from 1990 to 1992 and appear likely to increase again for 1994. In addition, early analyses of 1993 survey data, collected 3 to 4 months after the main release of mathematics results and before the release of the reading results, indicate that nearly half of the state assessment directors and mathematics specialists from participating states found the TSA to be a "very worthwhile" return for the effort; most of the remainder found it "somewhat worthwhile."

6. Evidence from surveys and case studies conducted in 1992 suggest that the 1990 TSA has had a positive impact on curriculum and assessment in many of the states that chose to participate, as well as some that did not. Among the states that were involved in curricular or instructional reform and had also participated in the 1990 TSA, 46 percent of the respondents indicated that changes were being influenced by NAEP and the TSA. The proportion was even larger among those states involved in *assessment* reform, where 51 percent of respondents from TSA states indicated an influence.

7. It appears that the 1990 TSA carried weight, not in isolation, however, but because it articulated well with other contemporary influences in mathematics education, particularly the NCTM Standards. In the Panel's judgment, the extent of this impact is likely to grow so long as state NAEP maintains its credibility as an "outside authority" that can be used to bolster the messages of state policy, curriculum, and assessment leaders.

The Panel's recommendations with regard to the use of achievement levels in reporting are presented in chapter 8 of this report. Additional recommendations for reporting are listed below:

- *NCES should establish and then adhere to a feasible and coherent schedule for reporting.*

- *In general, the speed of reporting should be increased and efforts should be made to avoid fragmenting the reporting of a given assessment. However, the Panel recognizes that partial reporting may still be the best alternative if it is the only way to avoid serious delays in the presentation of the main results.*

- *NCES should continue to seek input from NAEP's main constituencies regarding the formatting of the reports, and it should continue to investigate new and better ways to present various aspects of the data.*

- *Despite the fact that sustained, widespread media interest may be difficult to sustain over repeated cycles of the assessment, NCES should continue to hold press conferences to announce the main reports and otherwise continue to bring these materials to the attention of the press.*

- *NCES should also continue and expand upon the seminar format for presentation of the focused reports. In order to reach a wider audience, NCES should consider presenting the seminars at regional professional conferences or utilizing a video format.*

- *Additional efforts should be made to provide the states with materials suitable for use with district- and building-level personnel, including materials that translate state and national NAEP data into implications for instructional reform. The potential for using these materials to increase the willingness of schools to participate in the assessment should be examined.*

5 Guiding Principles

Much of the Panel's work has focused on technical issues related to the feasibility of a state NAEP, issues such as the adequacy of the sampling frame, participation rates, and possible performance differences as a function of session monitoring. However, an effective evaluation cannot be focused solely on narrowly defined technical issues. NAE has, therefore, continuously defined the role of the Panel also to include advice on policy matters that bear on the future of NAEP. In this chapter, the Panel articulates a set of principles for analyzing and evaluating these major policy issues. The remaining chapters in this section discuss each of the various policy issues in detail, and where appropriate, develop recommendations for action.

The National Indicator Principle

NAEP should continue to be the key indicator of what the nation's students know and can do, providing trend data on student academic performance in key subject areas.

As discussed in chapter 1, for nearly 25 years NAEP has been the key indicator of what the nation's children know and can do. Not only have NAEP data been useful in providing a snapshot of what the nation's children can do at a given point in time, but NAEP trend data also have been invaluable in monitoring educational progress across time. As such, NAEP has been a key indicator of the health of the nation's education system. In the words of the Alexander-James report:

> [NAEP] has become the primary source of information on what American students, nationwide, know and can do, covering a wide range of school subjects....Since by design that information is comparable in many respects from one assessment to the next, we can now compare the children of today with those of earlier years and learn how well our educational system is working. No other measure of educational achievement has NAEP's size, scope, and perspective over time.[1]

The quality of NAEP samples makes NAEP the most *accurate* national indicator of educational progress available. When college admission scores indicated that reading and mathematics achievement were dropping, with the strong implication that all student achievement was dropping as well, NAEP showed a more complex and accurate picture of student achievement. For example, NAEP scores showed that achievement in the areas of reading and (for 9- and 13-year-olds) science actually have been reasonably stable from 1969 to the present. NAEP also documented that science knowledge among 17-year-olds declined over the past 20 years, that mathematics proficiency among 9-year-olds increased over this same timespan, and that appreciable gains have been made in science and mathematics by Hispanic youth

[1] L. Alexander and H.T. James, *The Nation's Report Card: Improving the Assessment of Student Achievement* (Washington, D.C.: The National Academy of Education, 1987), 11.

and in mathematics by African-American students.[2] The value of these trend data for the nation is obvious. NAEP has played and should continue to play the same role for education that economic indicators such as the unemployment index play in documenting changes in the country's economic health. As innovations and improvements are made to NAEP over time, it is imperative to preserve NAEP's accuracy and integrity as an indicator.

The Lead and Reflect Principle

NAEP frameworks and content must be comprehensive, reflecting the most forward-looking pedagogical approaches, at the same time that they reflect the best of current practice.

The lead and reflect principle addresses the tension inherent in assessments that promote high standards while monitoring current practice. To be the national indicator of what students know and can do, NAEP must measure what students are currently taught. At the same time, to measure progress toward national goals and eventually the attainment of new, more ambitious content standards, the assessment must also include the full extent of these new content expectations. As the Panel stated in its evaluation of the 1990 TSA,

> NAEP must reflect a balance between innovation and proven pedagogy: Its frameworks must lead curriculum and instruction at the same time that they reflect the outcomes of current practice. If NAEP can maintain these balances, it will play an important role as a key indicator of the status of education and the needs for changes in policy and practice at the state and national levels.[3]

The Panel reinforced this position in its second report:

> NAEP frameworks should be broad enough to reflect goals for what students should know and be able to do as well as how they are presently being educated. NAEP content frameworks and test items should not shift wildly in response to fads within subject-matter disciplines, or NAEP will be unable to measure changes in learning.[4]

In short, NAEP must always be capable of measuring progress toward new levels of learning at the same time that it provides an accurate picture of what students are learning in the classroom.

[2] I.V.S. Mullis, J.A. Dossey, M.A. Foertsch, L.R. Jones, and C.A. Gentile, *Trends in Academic Progress* (Washington, D.C.: National Center for Education Statistics, 1991).

[3] The National Academy of Education, *Assessing Student Achievement in the States* (Stanford, CA: Author, 1992), 67.

[4] The National Academy of Education, *Setting Performance Standards for Student Achievement* (Stanford, CA: Author, 1993), 8.

The Policy Relevance Principle

NAEP must collect timely and relevant data for policymakers but not have its integrity corrupted by political pressures.

NAEP must not become a means for promoting partisan political agendas. Just as the government's labor indexes do not promote the Administration's labor policies, NAEP must continue to be an index of educational attainment, informative but policy neutral. Policy neutrality does not imply policy indifference, however; NAEP's primary role *is* the provision of relevant data for policy decisions affecting the nation's education system. As the Panel stated in its second report

> The Panel believes that NAEP itself should not function directly as an agent of education reform. If NAEP policy decisions are made to "leverage" education change, there is the risk that the indicator itself will become corrupted. If NAEP actively pursues a reform agenda, the content of the indicator will be dictated by the reform curriculum advocated at a given time and will shift erratically over time in response to changes in political authority. The Panel believes, however, that NAEP has an important role to play in education reform by providing data that inform the debate about needed changes.[5]

The State Indicator Principle

While NAEP's main role is that of the key nationwide indicator of educational progress, NAEP plays an additional needed role by providing information to the states on the educational progress of their students.

States, along with districts and local schools, play *the* key roles in setting education policies in the country. For this reason, the TSA was inaugurated to provide states with indicator data about the progress of their students in the same way that NAEP provides those data for the nation.

The Panel's survey of the states indicates that they find NAEP data a useful source of information on the educational performance of their students. The increase from 37 states participating in the 1990 TSA to 41 states in the 1992 TSA provides further evidence of the utility and interest that states see in state NAEP. Just as national NAEP has been useful in tracking changes in educational achievement across time, state NAEP will begin to inform states about trends in the educational progress of their

[5] Ibid., 149.

students. And, because NAEP results are reported for major subpopulations, states can compare their performances across various racial-ethnic and socioeconomic groups as well. Of course, state NAEP data also allow for comparisons of a state's results with those of the nation.

In summary, one of NAEP's roles is to provide to the states the same quality data that it provides for the nation. These data help inform states about the educational progress of their students in general, as well as important subgroups of their students.

The Inclusiveness Principle

To the degree technically, ethically, and financially possible, NAEP should assess an inclusive sample of all children whose ages would ordinarily place them in the 4th, 8th, or 12th grade.

Congruent with the idea of NAEP as a national indicator is the importance of inclusiveness, ensuring that NAEP reflects the proficiency of as many of the nation's children as possible. NAEP should include children in the three grades it has traditionally assessed—4, 8, and 12—and, insofar as possible, children who *would* be in those grades if they were in school.

As the timeline in chapter 1 indicates, NAEP assessed national samples of 9-, 13-, and 17-year-old students from its inception through 1988. In the 1984 assessment, NAEP expanded the samples to include students in grades 4, 8, and 11 as well. Beginning in 1990, new trend lines were begun for grades 4, 8, and 12, but bridge samples of 9-, 13-, and 17-year-olds were added to maintain the earlier trend lines. In earlier years, NAEP also included out-of-school 17-year-olds and young adults (ages 26-35), but the difficulty of sampling and data collection, combined with cuts in funding, eventually eliminated these groups from the sample.

NAEP excludes children deemed incapable of taking the assessment because of a serious disability (students with an individualized education plan [IEP]) or because of limited English proficiency (LEP). The expanding numbers of IEP and LEP children served by our nation's schools, the fact that schools composed entirely of children with disabilities are excluded from the NAEP sampling frame, and the fact that dropouts are not assessed all necessarily narrow the usefulness of NAEP as a broad indicator of what the nation's school-age children know and can do. Moreover, certain racial and minority groups for which NAEP provides separate estimates of achievement are disproportionately represented in special education populations.[6] For these reasons, insofar as technically, ethically, and financially possible, NAEP should be broadened to assess all children whose ages would ordinarily place them in the 4th, 8th, or 12th grade and who are capable of being assessed.

[6] National Council on Disability, *Serving the Nation's Students with Disabilities: Progress and Prospects* (Washington, D.C.: Author, 1993), 7.

The Relative Stability Principle

NAEP frameworks should remain unchanged for a reasonable period of time. The same framework must be in place long enough to establish meaningful trend lines, but not so long as to fail to accommodate the natural evolution that occurs in subject-matter fields.

NAEP's role in providing trend data demands that frameworks be stable enough to track meaningful changes in educational performance over time. However, as curriculum experts reach consensus about needed changes in a given subject-matter area, and as changes in the content standards for that area follow, the corresponding NAEP framework should change as well. Similarly, classroom practices change over time, and these changes also should be reflected in NAEP frameworks. Ordinarily, changes in both standards and practice are gradual, thereby allowing a framework to remain intact long enough to measure meaningful trends. In its second report, the Panel suggested a cycle of implementation, feedback, and revision that might take place over an 8- to 10-year period.[7]

The International Indicator Principle

Given the need for the United States to learn more effectively from educational practices in other nations and to compare practices and results, NAEP should, to the extent possible, be linked with major, relevant, large-scale international assessments.

Today's economy is increasingly international, sparking greater concern for the performance of our nation's children relative to those in other countries. Further, a useful cross-fertilization of ideas is engendered by comparisons of educational attainment and practices across countries.

While a number of formidable technical considerations confront those interested in international comparisons, including the equating of tests given in different languages and the creation of meaningful measures of opportunities to learn, these comparisons grow in importance as the world continues to shrink. Therefore, NAEP, as the pre-eminent indicator of the nation's educational progress, should be linked with major international assessments to the greatest extent possible.

The Assessment Innovation Principle

NAEP must continue to be the nation's leader in assessment innovation.

During its 25-year history, NAEP has been responsible for significant contributions to large-scale assessment, thus providing a research and development function that

[7] Ibid., 143.

benefits the entire assessment field. Examples include the development of sophisticated content frameworks, the use of innovative assessment procedures in various subjects, and the large-scale implementation of sample-based data-collection procedures that provide efficient estimates of student achievement for important population groups. NAEP has also been a leader in the use of powerful statistical tools that increase the precision of assessments.

Even greater innovation is likely to be required in future years than was required in the past. For example, while NAEP has begun to incorporate significant numbers of extended-response and other performance-based items in the assessment, the push continues for more items that tap advanced knowledge and higher order thinking skills. Even greater innovation will be required should NAEP be asked to meet the challenge of assessing at least some of the children currently excluded for reasons of limited English proficiency or physical or mental disability. New forms of NAEP may be required, including individually administered assessments for the learning disabled and Braille, large-type, foreign-language, and computer-adapted versions of the assessment. The challenge of devising such assessments is particularly daunting given the need to be able to compare their results to results from regular NAEP assessments.

Overview of the Remaining Chapters

The Panel now turns its attention to the major policy issues confronting NAEP, beginning in chapter 6 with the question most central to the Panel's evaluation: should a continuing state NAEP be authorized? Chapters 7 through 12 probe the issues of who should be assessed; how standards-based reporting can be implemented; whether to lift the prohibition against reporting NAEP results below the state level; how to articulate NAEP with emerging national content and performance standards; whether NAEP can be linked to state and international assessments; and how to maximize NAEP participation rates. The eight principles developed in this chapter are used to guide analyses of these issues, and where appropriate, to formulate recommendations for action. Finally, in the 13th and final chapter, the Panel considers the major design and policy trade-offs that confront NAEP in the face of current funding realities.

6 Should State NAEP Be Continued?

As Congress considers the reauthorization of NAEP, it must decide whether to change state NAEP from a trial to a full-fledged program. Of all the issues addressed in this report, this, of course, is the most important. In formulating a recommendation with respect to whether a permanent state NAEP should be authorized, the Panel considered three important questions: (1) Was the administration of the 1992 state NAEP technically feasible and sound? (2) Was the content of the 1992 mathematics and reading assessments valid? (3) Have state NAEP results been useful for the states?

The Technical Feasibility of State NAEP

In judging the technical feasibility and soundness of the 1992 TSA, the Panel considered the adequacy of the sampling design; school and student participation rates; session administration, including differences in the performance of students in the TSA versus the national assessment and in monitored versus unmonitored sessions of the TSA. State-by-state patterns of student exclusions based on individualized education plans (i.e., disabilities) or limited English proficiency were also explored.

The Sampling Design

The Panel concludes that the sampling design for the 1992 TSA was of high quality and was executed with care and attention to detail (as it was also in 1990). The requirement that the sample for each component of the assessment include about 100 schools and 2,500 students per state did, however, place a burden on small states where the sample size was close to the total number of available schools and, in some cases, exceeded the total number of available students.

Participation Rates

As detailed in chapter 2, the sampling frame, drawn by the NAEP contractor, informs the sampling of schools within a state. The NAEP contractor notifies the state coordinators about the schools in their state to be sampled, and the state coordinator then contacts the schools to elicit participation. The percentage of schools that agree to participate is called the *initial participation rate*. When schools refuse to participate, substitute schools also chosen by the contractor are then contacted by the

state coordinator for participation in the study. A state's *final participation rate* is the number of schools that were initially contacted and eventually participate plus those that are contacted as substitute schools and also participate.

On average, initial school participation rates for the 1992 TSA were adequate (89 percent for each grade), although significantly lower than the rate attained in the 1990 TSA (94 percent). However, the initial rates for the comparable grades in national NAEP in both years ranged from 86 percent to 89 percent, suggesting that the higher participation rate for the 1990 TSA might have been due to a "novelty effect." That is, interest in participation may have been unusually high because state NAEP was something new.

Compared to 1990, there also was increased *variability* between states in the initial school participation rates for 1992, and more states had before-substitution rates below 85 percent. In 1990, there were only two states with initial school participation rates below 85 percent; in the 1992 TSA, one-third of the states were in this category. The lowest 1992 participation rates were for Maine, which recruited only 62 percent of its originally sampled schools in grade 8 and 58 percent in grade 4.

For 1992, NAEP reports flagged results for states that had initial weighted school participation rates below 85 percent *and* final weighted school participation rates below 90 percent, as well as states that had initial weighted school participation rates below 70 percent. In addition, states could be flagged if the nonparticipating schools were concentrated within a particular class of schools, if the weighted *student* participation rate fell below 85 percent, or if nonparticipating students were concentrated within a particular class.[1] These flags appeared only in the appendix of the 1992 mathematics reports but were in the main text of the 1992 reading reports. For 1994, NAGB has passed a resolution indicating that states failing the overall participation-rate guidelines should be reported in a special section, separate from those states meeting or exceeding the participation criteria. In the Panel's view, the criteria used for noting states with low participation rates are too lenient. Further, the Panel agrees that the results for these states should be reported in a separate section, and not simply flagged. **Because there is a substantial likelihood of nonresponse bias when initially selected schools fail to participate, the Panel recommends that the results from states with initial school participation rates below 85 percent be reported separately from the results of states with initial rates of at least 85 percent, regardless of the state's final school participation rates.**

The potential seriousness of the low initial participation rates in some states is underscored by the Panel's finding that higher performance on the assessment was associated with lower initial participation rates. This finding was replicated in both the fourth- and eighth-grade samples, suggesting that the result was not due to chance. Furthermore, the Panel's analyses suggest that the finding was not due to one or two aberrant states. *Thus, there is a concern that states with low initial participation rates might have inflated results on NAEP, and the Panel finds some of the states' initial participation rates to be too low for accurate reporting of their 1992 TSA results.*

[1] For grade 4, five states were flagged because they failed to satisfy the guideline for overall participation rates, and one additional state was flagged because of under-representation of a class of schools. An overlapping group of six states was flagged at grade 8. No states were flagged for violating the student participation guidelines.

Because the analyses that led to these findings were performed only shortly before this report went to press, the Panel has not had the opportunity to undertake additional analyses that might help to explain the association between initial participation rates and states' achievement scores. The issue will be investigated further during 1994. In the meantime, the Panel is concerned that the credibility of the TSA could be jeopardized if this problem is not corrected. Therefore, **the Panel recommends that the contractor take steps to ensure that uniform procedures are followed by the state coordinators in recruiting schools to participate in state NAEP.**

Comparing the State NAEP Sample to a National NAEP Composite Sample

In its evaluation of the 1990 TSA, the Panel found a small but reliable difference between the mathematics performance of eighth graders in the TSA sample and a specially constructed matched subset of students in national NAEP. For the 1992 administration, no reliable differences between the three TSA samples and corresponding composite national samples were observed. Because the 1990 result suggested that there may have been differences in student motivation (e.g., students taking the TSA might be more highly motivated to do well) or differences in administration procedures (e.g., the use of local administrators in the TSA versus the use of administrators hired by the contractor in the national administration), the results from the 1992 administration were reassuring.

The Effect of Monitoring Sessions

In order to ensure the quality of the administration of the assessment at the local level, 50 percent of the sessions, chosen randomly, were monitored by quality control monitors. The results of the monitoring indicate that the administration of the assessment was adequate in the vast majority of sessions. Importantly, no differences in overall student performance were found between the monitored and unmonitored sessions in either the 1990 or 1992 TSA.

Exclusions Due to Limited English Proficiency and Individualized Education Plans

Overall, only about 2 to 3 percent of students sampled were excluded because of limited English proficiency; exclusions for LEP students ranged from less than 1 percent in some states to nearly 11 percent in California in the fourth-grade reading assessment. About 5 to 6 percent of the students sampled were excluded overall because of an individualized education plan. IEP exclusion rates varied from 2 to 7 percent across the states. Substantial consistency between the 1990 and 1992 TSA

exclusion rates was observed (e.g., the 1990 to 1992 correlation for IEP exclusions was +0.76), suggesting that the same states that were high (or low) in exclusion rates in 1990 tended to be high (or low) in 1992 as well. Between-state variations in LEP exclusions are to be expected, given the demographic differences among the states, but additional research is needed to understand why some states exclude more IEP students than others when there is no a priori reason to expect such differences. This issue is treated in more detail in chapter 7.

Based on its evaluation of sampling, participation, and administration procedures in the 1990 and 1992 trials, the Panel concludes that state NAEP is, most likely, technically feasible. However, more research is needed before a firm conclusion can be drawn. The Panel has identified two major concerns related to participation rates in its review of the technical feasibility of state NAEP. First, the negative association between initial school participation rates and average state assessment scores is potentially a serious matter; therefore, NCES and its contractor should immediately undertake research that examines this finding. Second, the current guideline for flagging the results of schools with low initial school participation rates is too lenient and needs to be made more stringent. The potential for serious nonresponse bias exists, creating the possibility that results for those states with low initial school participation rates could be biased.

The Content Validity of the 1992 Mathematics and Reading Assessments

For mathematics, the widely endorsed NCTM Curriculum Standards provide the definitive statement of new mathematics literacy. However, these standards were only available in draft form when the 1990/1992 NAEP Mathematics Framework was developed in 1988, and, as a result, neither the framework nor the items for the 1990 or 1992 assessments provided fully satisfactory coverage of mathematics as defined by these standards.

Although somewhat hampered by the existing framework, the quality of the mathematics items did improve between 1990 and 1992, particularly with the expansion of the size of the item pool and the inclusion of new item types better suited to measuring the mathematical competencies endorsed by the NCTM Standards. Nearly 60 percent of the 1992 fourth-grade items (as compared to 50 percent of the eighth-grade items in 1990) matched one of the four major themes of the standards in some nontrivial way. However, fewer than 10 percent of the 1992 fourth-grade items were judged to be exemplars of the standards. Furthermore, while the 1992 assessment contained more items in the upper reaches of the NAEP scale than the 1990 assessment, still more challenging items are needed.

The Panel's lead and reflect principle argues for NAEP content coverage that achieves a balance between emerging concepts of what students should know and the material actually taught in the classroom. ***In the Panel's judgment, the 1990/1992 Mathematics Framework and item pool still reflect too much of what is currently being taught and do not extend far enough to include content implied by the NCTM Standards. Despite incremental improvements, the fact that fewer than 10 percent of the 1992 items were***

judged as exemplars of the NCTM Standards provides evidence that the NAEP mathematics assessment needs to change considerably before it fully meets the Panel's lead and reflect principle. The 1994 Mathematics Framework is expected to provide substantial improvement in this respect.[2]

The 1992 reading assessment, based on a newly developed framework, is a significant improvement over previous NAEP reading assessments. The assessment makes excellent use of authentic passages and open-ended items. Although the reading community, unlike the mathematics community, has yet to coalesce around a single statement of standards, experts from the reading field generally viewed the framework as a reasonable balance between what would be ideal and what is currently being taught at the classroom level. However, more work appears to be needed to develop personal-response and critical-stance items broadly acceptable to the reading community. In addition, as was found to be true in mathematics, additional challenging items need to be added at the upper levels of the NAEP reading scale.

The Utility of State NAEP Results

The Panel's studies of the 1990 and 1992 mathematics assessments show that the states strongly support a state NAEP. Perhaps the most compelling evidence for this is the fact that the number of states participating in the TSA increased from 1990 to 1992 and appears likely to increase again for 1994. In addition, early analyses of the Panel's 1993 survey data, collected 3 to 4 months after the main release of the 1992 mathematics results, indicate that nearly half of the state assessment directors and mathematics specialists from participating states found the TSA to be a "very worthwhile" return for the effort, while nearly all of the remainder judged it "somewhat worthwhile."

Evidence from surveys and case studies conducted by the Panel in 1992 further suggests that the TSA in mathematics has had a positive impact on curriculum and assessment in many of the states that chose to participate. Because so many influences are operating at once, it is difficult to attribute changes solely, or even primarily, to the TSA. However, 46 percent of the respondents from states that were involved in curricular or instructional reform in mathematics *and* that had participated in the 1990 TSA, indicated that NAEP and the TSA influenced their reform efforts. Respondents credited the TSA with adding impetus to already planned or desired changes, accelerating discussions of state goals, and increasing pressure to align with national standards. Some very specific uses and influences attributed to the TSA included tipping the balance in favor of calculators in the classroom and providing sample items that could be used in teacher in-service sessions to illustrate emerging trends in mathematics education. Similar influences on assessment were cited in states undertaking reform in that area. For example, the TSA reportedly hastened the passage of accountability legislation, the adoption of new or revised state assessments, the use of calculators on assessments, and the design of assessment items.

[2] Although a 1994 framework was developed for mathematics, there will not be a 1994 mathematics assessment. The next mathematics assessment is now expected to occur in 1996.

Summary

Based on its evaluation of the 1990 and 1992 TSAs, the Panel summarizes its conclusions as follows:

- States are strong supporters of state NAEP and are using NAEP results to support ongoing education reform movements. Moreover, the number of states participating in 1992 was greater than in 1990.

- State NAEP appears to be technically feasible, and the sampling and administration procedures appear satisfactory. However, the negative relationship between initial school participation rates and overall state performance on state NAEP must be further explored. Because there is a substantial likelihood of nonresponse bias when initially selected schools fail to participate, the Panel recommends that the results from states with low initial school participation rates be reported separately from the results from states with acceptable initial school participation rates.

- The 1992 Reading Framework and items represent a reasonable balance between what content experts believe the nation's children should know and what is currently being taught in the classroom. The 1990/1992 Mathematics Framework and items, while moving in the direction of the NCTM Standards, do not yet represent an optimal balance between what mathematics curriculum experts say children should know and be able to do and what is currently taught in schools.

Based on its evaluation of the 1990 and 1992 trials, the Panel recommends that Congress authorize a continuation of state NAEP. However, there is a need for ongoing evaluation. Certain questions related to full-scale implementation of state NAEP have not been answered by the first two trials and will not be answered by the limited, one-grade, one-subject trial scheduled for 1994. Both the technical feasibility and the perceived burden to the states of conducting a state NAEP at three grade levels in multiple subjects remain to be evaluated. And, there has been no 12th-grade trial in any subject area. It may be more difficult to secure participation at the 12th-grade level because schools realize that 12th-graders' motivation to perform well on a test that does not affect them personally is often less than for lower grades. **For these reasons, the Panel recommends that Congress mandate an ongoing evaluation of state NAEP with regular feedback to Congress. Because many of the factors affecting the quality and feasibility of state NAEP are the same as those affecting national NAEP, the Panel notes that evaluation of the two programs could be integrated for maximum efficiency and effectiveness.**

7 Future of NAEP: Who Should Be Assessed?

Public Law 100-297, which authorized the 1990 and 1992 NAEP assessments, is silent on the issue of who should be assessed, except in its requirement that the assessment provide "a fair and accurate presentation of educational achievement" by using "sampling techniques that produce data that are representative on a national and regional basis... ."[1] Similar language is used to describe the state-by-state assessment. *The NAEP Guide* indicates that all sampled students should be assessed except for those LEP students and IEP students who are judged to be incapable of participating meaningfully in the assessment.[2] While NAGB appears to be guided by the NAEP authorizing legislation (P.L. 100-297) and *The NAEP Guide*, it does not have a separate policy statement about who should participate in NAEP.

The pending *Goals 2000* legislation is explicitly inclusive and argues for the development of challenging content and performance standards for *all* students. Given that NAEP is likely to have a central role in assessing the degree to which standards are being met, this is a propitious time to examine the NAEP policy on who should be assessed. If NAEP is to be an indicator of what the nation's children know and can do, then it should reflect, to the greatest degree possible, what *all* children know and can do. The inclusion of private school students beginning with the 1994 state assessment moves in the right direction. As noted in the Panel's evaluation of the 1990 trial, the exclusion of private school students from the trial assessments has made accurate state-by-state comparisons difficult (in part) because of large state-to-state variation in the proportion of students attending private schools.

In addition to private school student exclusions, the Panel's studies show, overall, that about 2-3 percent of the sampled students in the fourth and eighth grades were excluded because of limited English proficiency and another 5-6 percent because they had IEPs. These figures do not include the 4 percent of all 6-to-11-year-old and 8 percent of all 12-to-17-year-old IEP students who were not even sampled by NAEP because they were not enrolled in regular public schools.

Administering Special Versions of NAEP to Children with IEPs

The principle of inclusiveness would support including as many IEP students as possible in the NAEP assessment, and it seems probable that adapted, parallel assessments could be constructed and used effectively with many of these individuals. However, such a plan would be quite costly. For example, students with learning disabilities, who comprise a high percentage of excluded IEP students, are likely to require labor-intensive, individually administered assessments. Furthermore, some

[1] Public Law 100-297, Part C, Section 3403 6A: 1988.

[2] Educational Testing Service, *The NAEP Guide: A Description of the Content and Methods of the 1990 and 1992 Assessments* (Princeton, NJ: Author, 1990), 35.

students' disabilities are such that only partial testing would be possible, and, in a small number of cases, physical or mental disabilities are so severe that no type of assessment should be attempted.[3]

Prior to taking any steps toward assessing additional IEP students, the Panel believes that more research is needed to determine the severity of disabilities among those students enrolled in regular public schools who are presently excluded from NAEP. To do so, the Panel proposes to undertake two studies as part of the 1994 assessment.

The Panel has requested, and NCES has agreed to collect, comparable background data on all sampled IEP (and LEP) students *whether the student was excluded or not.* A modified version of the Excluded Student Questionnaire will be administered for IEP and LEP students included in the TSA. With these data, the Panel will be able to compare excluded versus included IEP and LEP students on variables such as type and degree of disability, or the percentage of time the student is mainstreamed or in a special language class. Second, the Panel will conduct onsite interviews with local assessment administrators in the states with high IEP exclusion rates and the states with low IEP exclusion rates in an effort to ascertain why the rates of exclusion differ across states. Third, and most important, the Panel will evaluate a sample of the students who have been excluded within these states to form an independent judgment of their "assessability." That is, the students' disabilities will be specified, and a determination will be made from among the following possibilities: the student could have been assessed with the regular NAEP assessment; the student could have been assessed with an alternative or adapted assessment and, if so, which type; or the student could not have been assessed at all.

In summary, the Panel expects to determine, based on these data, (1) the degree to which the criteria for exclusion are being applied consistently from state to state, (2) whether additional students could or should have been assessed, and (3) the types of assessments required to assess students currently being excluded because of a disability. The Panel will report on the results of its research on the exclusion of IEP and LEP students and make a recommendation based on that research.

Administering Special Versions of NAEP to Children with LEP

In *Assessing Student Achievement in the States,* the Panel took the position that serious consideration should be given to the development of non-English versions of NAEP. Only about 2 percent of eighth graders and 3 percent of fourth graders were excluded from the 1992 TSA for reasons of limited English proficiency. The Panel noted significant state-by-state variation in exclusion rates, however, with LEP exclusion rates ranging from less than 1 percent in each grade in Alabama to about 5 percent of the eighth graders and 10-11 percent of the fourth graders in California. While the difference between fourth- and eighth-grade LEP exclusion rates seems somewhat small, the differences by state are not surprising given the ethnic compositions of the states. Nevertheless, these state-by-state variations do impact the ability to make valid cross-state comparisons using NAEP data.

[3] J.M. Kaufmann, "How We Might Achieve the Radical Reform of Special Education," *Exceptional Children* 60 (1993): 6-16.

When considering the issue of non-English versions of NAEP, a concern often raised is the number of different languages spoken in the United States. The 1990 Census found that nearly 32 million persons in the United States speak a language other than English (about 14 percent of the population).

Without question, NAEP could never be expected to generate assessment instruments in more than a handful of the languages most commonly spoken in the United States. As figure 7.1 shows, Spanish is by far the most common language spoken in the United States (after English). Furthermore, the number of persons from Spanish-speaking backgrounds has increased from about 6 to 9 percent of the nation's population between 1980 and 1990. Roughly 22 million Americans are of Hispanic background.[4] Therefore, if a non-English version of NAEP is to be constructed, it should be constructed in Spanish. To help make a decision about whether or not to recommend the development of a Spanish-language version of NAEP, the Panel commissioned a paper examining the pros and cons of doing so.[5]

The most compelling reason for establishing a Spanish-language version of NAEP is to provide an accurate assessment of the achievement of the nation's Spanish-speaking LEP population. Establishing a Spanish-language version of NAEP would provide students with the opportunity to be tested in subjects such as mathematics and science in the language they believe would most effectively demonstrate their achievement in these subjects. Furthermore, a Spanish-language NAEP would allow Puerto Rico to participate in the TSA. A precedent has been set for Spanish-language achievement testing in that several local school districts, and some states already offer tests in Spanish to elementary students who have been excluded from their regular assessment programs.

Several negative arguments also can be advanced against the development of a Spanish-language NAEP. First, difficult technical problems would have to be overcome. Among others, there are questions about the scoring comparability of English and Spanish constructed-response items; about cultural differences affecting the translation of items; about differences between Puerto Rican, Mexican, Cuban, and other Latin American dialects; and about the feasibility of equating Spanish NAEP scores to English NAEP scores. Further, the establishment of a Spanish-language version of NAEP would be costly and would still not address the exclusion of other language-minority groups. Ultimately, criteria would have to be developed to stipulate the population percentage at which a language-minority group would qualify for its own version of NAEP.

Thus, establishing a non-English version of NAEP would have substantial financial and political ramifications above and beyond the technical issues. As suggested above, there are strong reasons both for and against establishing a Spanish-language version of NAEP. Based on the principle of inclusiveness, *the Panel supports the development of a research agenda, including pilot studies, to examine the feasibility of developing a Spanish-language version of NAEP.*

[4] U.S. Department of Commerce, "Number of Non-English Speaking Americans Up Sharply in the 1980s, Census Bureau Says," *U.S. Department of Commerce News* (April 1993).

[5] W.G. Secada, "Issues in the Development of Spanish-Language Versions of the National Assessment of Educational Progress," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

Figure 7.1. ***The top 25 languages***

The total number of United States residents 5 years old and over speaking a language other than English at home in 1990 was 31,845,000 (14 percent of the 230,446,000 residents 5 years old and over). This is an increase of almost 38 percent over the 1980 total of 23,060,000.

Language used at home	Total speakers over 5 years old		Percentage change
	1990	1980	
Spanish	17,339,000	11,549,000	50.1
French	1,703,000	1,572,000	8.3
German	1,547,000	1,607,000	-3.7
Italian	1,309,000	1,633,000	-19.9
Chinese	1,249,000	632,000	97.7
Tagalog	843,000	452,000	86.6
Polish	723,000	826,000	-12.4
Korean	626,000	276,000	127.2
Vietnamese	507,000	203,000	149.5
Portugese	430,000	361,000	19.0
Japanese	428,000	342,000	25.0
Greek	388,000	410,000	-5.4
Arabic	355,000	227,000	57.4
Hindi, Urdu, and related	331,000	130,000	155.1
Russian	242,000	175,000	38.5
Yiddish	213,000	320,000	-33.5
Thai	206,000	89,000	131.6
Persian	202,000	109,000	84.7
French Creole	188,000	25,000	654.1
Armenian	150,000	102,000	46.3
Navajo	149,000	123,000	20.6
Hungarian	148,000	180,000	-17.9
Hebrew	144,000	99,000	45.5
Dutch	143,000	146,000	-2.6
Mon-Khmer	127,000	16,000	676.3

SOURCE: Census Bureau
Calculations are from numbers before rounding

Puerto Rico received a National Science Foundation grant to investigate the development of a Spanish-language version of NAEP for the 1994 assessment and, with the permission of NCES, has contracted with the Educational Testing Service to develop it. The assessment will be in two subjects—mathematics (using the 1990/1992 Mathematics Framework) and science (using the 1990 Science Framework). To examine the feasibility of a Spanish-language version of NAEP, the Panel therefore suggests two steps. First, the experience with the administration of a Spanish-language version of NAEP in Puerto Rico should be studied and closely monitored by NCES. Second, additional research should be conducted to document the experience of administering a Spanish-language NAEP to students on the U.S. mainland. This additional research is necessary for the following reasons. Most instruction in Puerto Rico is in Spanish, which is the primary language spoken in the home and with peers. In the United States, most instruction is in English, even for those students who speak Spanish exclusively with parents and peers and who are in bilingual programs. Furthermore, if a single Spanish-language version of NAEP is administered in the United States, not only students with a Puerto Rican background, but students with Mexican, Cuban, and Central and South American backgrounds would take this common version.

The Puerto Rico experiment will be useful for determining whether it is possible to translate NAEP into Spanish and maintain the original meaning of the items. This determination can be made through back translation of the translated material and by an examination of the item-characteristic curves. However, for the reasons mentioned in the previous paragraph, these analyses will not be sufficient to determine whether a full-scale Spanish-language version of NAEP is feasible, nor will they help determine whether performance of U.S. Spanish-speaking students on NAEP is improved by having a Spanish-language version of the assessment available. **The Panel recommends that as part of the 1996 national assessment, NAGB authorize a set of pilot studies in locales where a significant proportion of students speak Spanish as their first language (e.g., Arizona, California, Florida, Illinois, New Mexico, New York). The pilot studies should include an experimental design that allows NCES to determine whether NAEP results vary as a function of administering a Spanish-language version of NAEP for Spanish-speaking students.** For example, the Panel suggests that three versions of the assessment be constructed and randomly assigned to Spanish-speaking LEP students. The first version would be in English, the second would be in Spanish, and the third would provide English and Spanish versions side-by-side so that the student could select the language in which he or she expects to most effectively demonstrate achievement. In analyzing results from this study, it will be important to examine the effects of the degree to which Spanish is spoken at home, whether some or all peers speak Spanish exclusively, how much instruction the student receives in Spanish and how much in English, and whether the student is in a bilingual program.

An important issue is whether all or only some subjects should be assessed in Spanish, assuming a Spanish-language version of NAEP is proven feasible. *The Panel has considered this matter and recommends that the pilot Spanish-language versions of NAEP should be in mathematics and science, two subjects where understanding of the subject matter is not integrally dependent on the language spoken.* If the mathematics and science assessments prove feasible, the next step might be to develop pilot

Spanish-language versions of the reading and writing assessments. Since literacy is a skill not easily decontextualized from the language in which one is literate, the aim of the latter would *not* be to equate the Spanish assessments with the English reading and writing assessments, but rather to assess Spanish literacy as a valued outcome in itself.

Administering NAEP to Out-of-School 17-Year-Olds

In *Assessing Student Achievement in the States,* the Panel recommended that state NAEP reflect the achievement of all 4th-, 8th-, and 12th-grade students. The Panel noted that NAEP is presently an assessment of *enrolled* 4th, 8th, and 12th graders and pointed out that a significant proportion of children fail to complete their schooling in 12 years. Furthermore, the Panel noted that students who drop out of school are likely to score lower on standardized tests, have lower grade-point averages, and are disproportionately more likely to be Hispanic or African-American. The Panel also noted that the composition of the dropout group has changed and continues to change, calling into question the comparability of 12th-grade NAEP results over time.

Perhaps of even greater importance is the paucity of information about the skills and knowledge of the dropouts. While the recently completed adult-literacy study included a stratum of 16-to-18-year-olds, the results are not broken out by whether the respondents are in or out of school. Even if that information had been provided, the measures used in the survey are very different from the NAEP measures.[6]

In its report evaluating the 1990 TSA, the Panel recommended that beginning with the 1994 assessment, a national sample of out-of-school 17-year-olds be assessed. To the Panel's knowledge, this recommendation was never discussed by NAGB or NCES. Following from its principle of inclusiveness, the Panel continues to believe that it is important to learn what the nation's out-of-school 17-year-olds know and can do. However, the Panel recognizes that identifying and testing this sample could be prohibitively expensive. **If Congress fails to provide sufficient funding for NAEP to assess out-of-school 17-year-olds, the Panel recommends that future adult literacy surveys oversample 16-to-18-year-olds, stratifying on whether they are in or out of school, in order to provide the nation with a better picture of the literacy of the nation's out-of-school youth. Further, selected blocks of NAEP items should be included on the adult literacy survey to increase comparability between the two assessments. Finally, given that the proportions of out-of-school 17-year-olds vary across time and across the states, the Panel recommends that all 12th-grade state NAEP results released in the future also include information on the percentage of 17-year-olds who have not completed high school and that these percentages be reported alongside each state's results.**

[6] I.S. Kirsch, A. Jungeblut, L. Jenkins, and A. Kolstad, *Adult Literacy in America* (Princeton, NJ: Educational Testing Service, September 1993).

8 How Should NAEP Results Be Reported?

NAEP results must be reported in ways that are both technically defensible and easily interpretable, as well as useful to policymakers and the public. In 1989, in the context of the Charlottesville Education Summit and the development of national education goals, NAGB decided to establish achievement levels—that is, performance standards on the NAEP score scale that specify cutscores that students in a given grade must reach to be considered advanced, proficient, or basic. The level-setting effort was predicated on the assumption that assessment results would be more meaningful if actual student performance was reported in comparison to substantive standards specifying what students *should* know. Setting achievement levels on NAEP was seen as a way to participate in the national agenda for setting educational goals and to enhance the usefulness of NAEP for measuring progress toward the goals.

The Panel's Evaluation of the Achievement Levels

Due to the salience of the standards issue and controversy surrounding evaluations of the 1990 achievement levels, NCES asked the NAE Panel to expand its evaluation of the 1992 TSA to include an extensive study of the 1992 achievement levels. As described in its report, *Setting Performance Standards for Student Achievement*, the Panel found serious problems with the methodology used to set the 1992 achievement levels and with the resulting cutscores.[1] ***The Panel recommended that NAGB discontinue the use of the Angoff method to set achievement levels, discontinue reporting using the flawed 1992 levels, and not reinstate achievement levels as the primary reporting mechanism until the following conditions are met: (1) a new, more defensible way of setting the achievement levels can be developed, (2) the achievement levels have been empirically evaluated prior to using them as the primary NAEP reporting mechanism, and (3) the assessment content is better articulated with emerging national content and performance standards.***

The Angoff Method Is Fundamentally Flawed

Based on a series of internal analyses, the Panel concluded that the Angoff method for setting standards is fundamentally flawed. In both reading and mathematics at every grade level, the findings were the same. Judges who set the 1992 levels were unable to make consistent judgments when translating their substantive expectations into cutscores on the NAEP scale. In some cases, internal inconsistencies in judges' boundary scores intended to define a standard were as large as four to eight grade levels on the NAEP scale. Because inconsistencies occurred in response to item type (such as multiple-choice versus short-answer questions) and not substantive

[1] The National Academy of Education, *Setting Performance Standards for Student Achievement* (Stanford, CA: Author, 1993).

dimensions of the assessment, the problem was attributed to the standard-setting method itself. To implement the Angoff procedure, judges must imagine hypothetical groups of students performing just at the lower boundary of each of the intended achievement levels. Judges then go through all the questions in the assessment and estimate the proportion of students in each boundary group who will answer each question correctly. The large internal inconsistencies that resulted from this process suggest that item-by-item difficulty judgments are not a good way for judges to arrive at a substantive standard.

The Panel Endorses Standards-Based Reporting

While criticizing the methods used to set the 1992 achievement levels, NAE Panel members strongly endorsed the potential value of reporting by standards that exemplify challenging curricular and performance expectations. As part of its evaluation studies, the Panel collected validity evidence using alternative methods for setting standards and made recommendations about how these approaches, with modifications, could be used to develop achievement levels for the 1994 assessments. *The Panel recommended that at a minimum, all three of the following approaches be used: (1) contrasting-groups field-based studies, (2) an item-mapping procedure, and (3) a total-student-performance (whole-booklet) procedure.*

The contrasting-groups method requires that students be identified for whom it is confidently concluded that they meet the substantive standards for advanced, proficient, and basic performance. In the case of the Panel's studies, this was done by a combination of teacher nominations and researchers' individual assessments. Then an assessment-score standard is set by finding the score that best distinguishes advanced from proficient students, and so forth. The item-mapping procedure involves locating (or mapping) all the assessment items onto the score scale and then asking judges to deliberate about where cutscores should be set by comparing items in each region of the scale to the substantive standards. The whole-booklet procedure is analogous to the contrasting-groups method except that actual student responses in assessment booklets are used by judges to identify students who clearly meet the definition for advanced, proficient, and basic. Then judges can deliberate directly about where the cutscore should be set to best distinguish these groups. The advantage of the whole-booklet approach over the item-by-item difficulty method is that it allows judges to consider more integrated examples of performance and agree on patterns of right and wrong answers that exemplify advanced or proficient performance.

New Achievement Levels Must Be Treated as a Research and Development Effort and Be Rigorously Evaluated

The Panel emphasized that achievement levels developed using these approaches or any other method should be rigorously evaluated before becoming an operational part

of the reporting of NAEP results. It is customary for statistical agencies to try out new indices and establish their accuracy, validity, and interpretability prior to adopting them as the regular means of reporting. Therefore, in its previous report, ***the Panel recommended that achievement levels developed for 1994 should be reported in parallel with the official NAEP reports and identified as draft or developmental standards.*** Tentative or experimental reporting of achievement levels would avoid previous problems that have made it necessary to revise the achievement levels after the fact and recompute trend data. In addition, it would allow for professional critique and revision of these initial levels, as well as ensure that NAEP frameworks have time to be aligned as much as possible with content standards now being developed in each subject area, before the achievement levels become permanent. Because valid assessment of trends over time requires holding standards constant for a reasonable period of time, official achievement levels in a subject area should not be adopted until these conditions are met.

The Panel's recommendation to delay using achievement levels as the primary means for reporting until their validity can be evaluated implies that in the interim, NAEP results should be reported by traditional methods—mean scores and descriptive anchor points. The Panel has been criticized for not subjecting anchor points to the same degree of scrutiny as has been applied to the achievement levels. However, it is the Panel's view that anchor points are merely descriptive of the NAEP scale and do not claim to be standards for what students *should* know. Because anchor points claim less and do not invite such strong interpretations as suggested by achievement levels, they do not have to meet all of the same validity criteria. Moreover, the anchor points have the following technical features that make their descriptive accuracy less problematic. (1) There is no dispute about where the anchor points should be because they are specified in standard deviation units. (2) Exemplar items are located using statistical criteria. (3) Anchor points are described using narrative phrases that are closely tied to assessment content and therefore more accurately convey how students at a given location on the scale performed on the assessment.

Achievement Levels Require That NAEP Content and Performance Standards Be Consistent with National Standards

The use of achievement levels to communicate standards to the public and to measure progress toward national goals places new, much more stringent demands on the substantive content of the assessment. The Panel's content validity studies documented that the NAEP item pools in reading and mathematics have improved significantly. Nonetheless, the 1992 item pools, particularly for items that serve to distinguish advanced students, are not sufficiently congruent with emerging national content standards to communicate these content standards to the public. The quality of released items is especially troublesome. The Panel's content experts were concerned that reporting achievement levels that reify a traditional skills-before-thinking kind of curriculum could do more harm than good.

Future development of achievement levels requires that NAEP content and performance standards be consistent with national standards. NAEP cannot measure less than is implied by new content standards or it will be incapable of measuring

progress toward their attainment. However, in accord with the Panel's earlier recommendation that NAEP be comprehensive with respect to competing views of curriculum and current education practice, NAEP also must have a more inclusive content domain than implied by future-oriented, nationally certified content standards. For example, NAEP should continue to assess currently operational curricular goals in addition to those implied by new content standards as a way to measure change over time. As the content of the assessment is expanded to represent the full extent of ambitious standards in mathematics and reading, the Panel advised that greater attention be given to the reasonableness of combining old and new content into a single composite score. An alternative reporting practice would be to report separate trend lines for a "current-practice assessment" and a "national-standards-aligned assessment." The need for NAEP frameworks to articulate with emerging national standards is treated more fully in chapter 10.

International Comparisons Should Be Made

In addition to its recommendations about how to develop more defensible performance standards, ***the Panel also recommended that whenever possible, international comparisons should be provided to help evaluate the performance of U.S. students***. For example, the score corresponding to the 90th percentile of the four highest scoring countries on an international assessment (such as the upcoming Third International Mathematics and Science Study) could be identified as "internationally advanced," and the percentage of U.S. students reaching this level could be determined. The Panel cautioned that international comparisons do not help establish the validity of substantive standards in relation to the NAEP scale, but international comparisons are interpretable and important in their own right.

Move to Within-Grade Score Reporting

At present, NAEP performance for students in grades 4, 8, and 12 is reported on a single 0-to-500-point scale. Furthermore, because of the way that the score scale was constructed, actual student performance is effectively compressed into a range of only 200 points (150 to 350). As part of its argument for achievement levels, NAGB expressed concern that use of a single scale obscured variation within grade and made evaluative interpretations more difficult.

The Panel considers the use of a single cross-grade scale to be problematic on other grounds. In its first report, the Panel noted that reporting a single composite score in mathematics was technically misleading and would not provide information useful for improving mathematics instruction in specific content areas such as geometry, algebra, measurement, and statistics. The decision to have a subject-matter scale span multiple grades further complicates the meaning of a composite score because weighting of the content areas in the composite changes with grade level. Even within content areas, the assumptions of the scaling methodology are strained by combining performance

across eight grade levels. Although the Panel does not have definitive evidence, a comparison of mathematics items mapped to the NAEP scale suggests that equivalent scores do not have the same meaning across grades.[2, 3]

During the course of its evaluation studies, the Panel heard many comments from subject-matter experts proclaiming that a strength of the achievement levels was their focus on describing performance within grade level. The Panel also believes that within-grade scales will make it easier for subject-matter experts to comment meaningfully on the quality of NAEP performance. For these reasons, *the Panel recommended moving to within-grade scaling effective with the 1994 assessment.*

In summary, the Panel concluded that the methodology of achievement levels and the resulting cutscores used to report NAEP results in 1990 and 1992 were fundamentally flawed due to substantial internal inconsistencies in the standard-setting process. Based on validity evidence from NAE evaluation studies, the Panel recommended that in future assessments three alternative approaches be used to set standards: (1) contrasting-groups field-based studies, (2) an item-mapping procedure, and (3) a total-student-performance (whole-booklet) procedure. In addition, the Panel recommended that any new achievement levels be rigorously evaluated before being established as a regular means of reporting results. With regard to NAEP content, the Panel recognized improvement in the quality of the NAEP item pool. However, NAEP's role as a comprehensive monitor requires that additional effort be made to align the content of the assessment with national content standards. Especially, reporting by achievement levels requires that the assessment include content adequately representing advanced performance. The Panel also recommended that international comparisons be provided along with achievement levels as another means to interpret student performance. Finally, the Panel strongly supported a move toward within-grade score reporting, effective in 1994, instead of using a single 500-point scale to report results across all grade levels. This would make the score scale more useful to educators and more understandable to the general public.

[2] G.W. Phillips et al., *Interpreting NAEP Scales* (Washington, D.C.: National Center for Education Statistics, 1993).

[3] The Panel is primarily concerned with the equivalence of scores across grades for the composite scale. It is aware that the appropriateness of the vertical scale is checked within content area by verifying that cross-grade items scale similarly for students in different grades.

9 Should the Prohibition Be Lifted?

Currently, reporting NAEP data below the state level is prohibited by law. As part of its evaluation of the 1990 Trial State Assessment (TSA), the Panel examined whether this prohibition should be removed and concluded that the ban should not be lifted. In spite of numerous and good reasons for lifting the ban, in the Panel's view, the risk of corrupting the national assessment is too high. Specifically, the Panel concluded that lifting the ban would increase the stakes for doing well on NAEP and therefore almost certainly lead to teaching to the assessment. The Panel cited research evidence suggesting that teaching to tests invalidates the results of the tests in that these results eventually will not generalize to other tests assessing knowledge in the same content areas.[1]

Continuing Calls for Lifting the Prohibition

Within the past year, NAGB in its *Positions on the Future of National Assessment Educational Progress* paper and the Department of Education in its draft legislation for the reauthorization of NAEP have both called for lifting the ban.[2,3] Both favor allowing districts and schools to use and report the results of NAEP data but with the proviso that they bear the costs of data collection, analysis, and reporting.[4] Under neither plan could data be reported at the individual student level.

In the Panel's view, it is inappropriate to administer NAEP to entities too small to support the current (state) NAEP design. This design involves sampling of both students and items. In this manner, a very large item pool is assessed at each grade level, allowing unusually thorough coverage of the subject area with relatively brief testing time per student. Individual schools do *not* have sufficient students to implement the design; modifying the design would have serious implications for the scope of the assessment. Narrowing the assessment to make it feasible to obtain complete and reliable assessment results for a single school would in turn make the assessment more vulnerable to efforts to teach to the test. For very large districts, however, it is feasible to propose implementing the NAEP sampling design as it has been applied to the states. In fact, the District of Columbia, a single school district, has already been included in state NAEP, and a number of the largest districts have more students than the least populated states.

[1] D.M. Koretz, R.L. Linn, S.B. Dunbar, and L.A. Shepard, "The Effects of High-Stakes Testing on Achievement: Preliminary Findings About Generalization Across Tests" (paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, April 1991), 1-30.

[2] National Assessment Governing Board, *Positions on the Future of National Assessment of Educational Progress, 1993* (Washington, D.C.: Author, 1993).

[3] *Report on Educational Research*, August 18, 1993, 3.

[4] According to NCES, the districts or schools could either get the raw data and analyze it themselves, or they could pay the NAEP contractor to analyze the data and to prepare the reports.

Given the interest expressed by both the Governing Board and by the Department of Education, the Panel therefore decided to revisit the issue of lifting the ban against reporting data at the *district* level.

Arguments in Favor of Lifting the Ban

Reasons for lifting the ban include the following:

- For those districts large enough to accommodate the NAEP sampling requirement, the availability of data at the district level would allow for the vertical integration of educational performance data. That is, comparable performance data would be available at the national, state, and district levels.

- Large districts would have access to an assessment instrument whose frameworks are based on state-of-the-art research and theory in the subject areas assessed. Furthermore, the frameworks represent a consensus about what all students should know.

- The incentive for participation in NAEP assessments might be greater. As noted in chapter 2, the increased burden that state NAEP has created may be making it more difficult to secure district- and school-level cooperation. Providing some participating districts with feedback about their performance would almost certainly increase interest in NAEP and thereby encourage participation.

- State NAEP appears to have increased the interest in and support for NAEP results more generally. It is likely that reporting for large-city school districts would further increase support for NAEP, while local media would almost certainly report on NAEP results in areas where district NAEP results are released.

Arguments Against Lifting the Prohibition

As good as these reasons are for lifting the ban for large districts, the Panel finds the reasons for maintaining the prohibition more compelling:

- If NAEP data were available at the district level, they would almost certainly be used for accountability; that is, NAEP would become a high-stakes assessment. The higher the stakes of an assessment, the more likely teachers are to teach to it. Research shows that teaching to a test will eventually erode the usefulness of the test to provide reliable trend information; performance on

that test may not generalize to the content areas being assessed. The generalizability problem exists no matter how well a test reflects the principles of good pedagogy.[5] Finally, as Beaton points out, the higher the stakes of the test, the more likely that "boosterism" will occur. Intensive efforts to encourage students to do well on the assessment could introduce bias in district-level results.[6]

- Other innovative and influential assessment initiatives might be discouraged. If NAEP, with state-of-the-art frameworks and national comparison data, were available for district-accountability purposes, states might no longer be willing to invest in the development of innovative assessment programs. Yet, in the past, such programs have contributed to the state of the art in assessment technology.

- While district-level assessment appears to be technically feasible, it would be complicated and expensive to implement.[7] Furthermore, important caveats must be noted. First, the design assumes that the measurement, administration, scoring, and reporting are the same as is done for state NAEP. Second, the sampling plans that could be employed all require high levels of school- and student-level participation in order to guarantee acceptable standard errors for estimates of performance. Third, and most important, the degree of bias that the higher stakes of a district-level assessment would introduce is unknown.

- Item security could become a problem. The design of NAEP depends heavily on the re-use of items. Concerns about security would require many more items to be replaced after each assessment cycle. This would lead to an increase in time required for item development and in the cost of the assessment, and it could affect the ability of NAEP to produce valid trend lines.

- If a district-level NAEP were to be mounted, another powerful constituency would be formed, putting new and different demands on NAEP. The states, by virtue of their participation, have become a potent constituency with considerable influence on the course of state NAEP and, indirectly, the entire NAEP program. Unlike the states and districts, however, national NAEP has no natural constituency to protect its integrity. It has few voices to argue against changes that could affect its usefulness as the key national monitor of educational progress.

[5] "Teaching to a test" does not mean simply emphasizing the principles and concepts embodied in the assessment, but rather, practicing on parallel exercises. Substantial evidence exists to confirm that this happens for the standardized tests commonly in use today.

[6] A.E. Beaton, "Methodological Issues in Reporting NAEP Results at District and School Levels" (paper prepared for the National Assessment Governing Board, October 2, 1992), 6.

[7] Ibid., 14.

To the Panel's knowledge, no data exist indicating the level of interest that districts have in participating. *Prior to lifting the ban, it would be prudent to determine how many districts would be willing to pay the cost for collecting, analyzing, and reporting NAEP data.* Given the likely cost of participation, the Panel suspects only the largest districts could even entertain the possibility of participation.

In summary, the Panel understands why some districts and schools might like to use NAEP to assess their students. State NAEP appears likely to become a powerful tool for the states in monitoring the educational progress of their students. Districts and schools would no doubt find NAEP trend data useful for the same purpose. The prospect of being able to link school and district data to the key national and international indicators of educational progress is a second reason. As legitimate and compelling as these reasons are, the Panel continues to believe that the most important of its principles is the key indicator principle. That is, every effort must be made to guarantee that NAEP be preserved as a monitor of educational progress at the *national* level. If NAEP were to become corrupted, it would lose the very features that now make it attractive to the states, districts, and schools. As the Panel stated in *Assessing Student Achievement in the States,*

> NAEP has been and should continue to be a *global* monitor of educational achievement. Its historic role as an independent auditor that neither is unduly influenced by particular instructional or testing practices nor shapes day-to-day classroom activities would be lost if TSA results were to be reported at the district, school, or student levels. In addition to compromising NAEP's value as an indicator, lifting the prohibition could detract from local innovations that attempt to integrate assessments with curricula.[8]

The Panel believes this statement to be as true now as it was when written, and therefore, **the Panel recommends that Congress continue the prohibition against reporting NAEP data below the state level.[9] If Congress weighs and reads the evidence presented in this report and decides to lift the ban, the Panel recommends that (1) the ban against reporting data at the *school level* remain in place, and (2) that district-level NAEP be implemented *on a trial basis* subject to the following four conditions. First, the same objectives, booklets, and manipulatives as are used in national NAEP would have to be used in district-level NAEP. Second, the sampling plan used would have to meet appropriate statistical requirements. Third, administration procedures for district-level NAEP would have to be identical to those used in state NAEP. Fourth, districts that participate should be required to help fund and participate in a Congressionally mandated independent evaluation.** The independent evaluation should be designed to address the issues the Panel noted above. In addition to examining the positive uses of assessment information, the evaluation should collect evidence relative to whether (1) teachers are teaching to the NAEP assessment, (2) assessment innovations at the state level are being eroded, and (3) districts are having an unduly large impact on NAEP policy decisions.

[8] The National Academy of Education, *Assessing Student Achievement in the States* (Stanford, CA: Author, 1992), 64.

[9] Panel member Gordon Ambach continues to dissent from this position; he is on record elsewhere as recommending the lifting of the prohibition at the district level where the size of the enrollment allows sampling as used at the state level. Panel member Edward Roeber dissents as well.

10 How Should NAEP Frameworks Be Related to Emerging National Content Standards?

Groups as diverse as NCTM and the National Governors Association have focused national attention on the issue of standards for instructional and assessment content. The pending *Goals 2000: Educate America Act* would create a National Education Standards and Improvement Council which, along with the National Education Goals Panel, would be responsible for certifying national content standards in a wide range of subjects. To remain credible, NAEP frameworks and items must be informed by these emerging national content standards.

Influence of the Standards Movement

The process of change and evolution in NAEP content frameworks and items has been ongoing since NAEP's inception, and alignment with national content standards has already begun. The current movement toward content standards was largely stimulated by the development of mathematics content standards by NCTM,[1] and these NCTM Standards were available in draft form when the objectives for the 1990 NAEP mathematics assessment were being developed. As a result, the 1990/1992 NAEP Mathematics Framework shows some evidence of influence from the NCTM Standards. For example, new types of items measuring higher order thinking skills in mathematics were added to the 1990 NAEP assessment and again to the 1992 assessment.

Also in response to the concern that the NAEP instruments be capable of measuring progress toward national standards, the use of extended open-ended items that began with the 1992 NAEP assessments in reading and mathematics will continue for all subject areas in 1994. In the view of the Panel, this trend is highly desirable and should be encouraged. It does, however, raise new technical and cost issues that deserve thoughtful consideration. For example, constructed-response items have required major innovations in scoring and scaling and have also proven quite costly to score. Also, constructed-response items have repeatedly been found to have lower response rates on NAEP than items with less demanding response formats.[2] Whether this is a matter of student motivation, a paucity of student strategies for approaching open-ended problems, time constraints on the administration of NAEP tests, placement of extended constructed-response items (usually at the end of the item block), or other factors is not known. *The Panel strongly suggests that NCES ask its Technical Review Panel to conduct research on this important issue.*

[1] National Council of Teachers of Mathematics, *Curriculum and Evaluation Standards for School Mathematics* (Reston, VA: Author, 1989).

[2] D. Koretz, T. Skewes-Cox, and L. Burstein, "Omitted and Not-Reached Items in Mathematics in the 1990 National Assessment of Educational Progress" (University of California at Los Angeles: Center for Research on Evaluation, Standards, and Student Testing, September 1992).

NAEP Frameworks Must Lead and Reflect

As discussed in chapter 5, the Panel's lead and reflect principle requires a balance between new content and new item types and those encountered in everyday classroom instruction. If NAEP assessments change more rapidly than instructional practice, NAEP will increasingly tap material and skills that students have not had the opportunity to master. This tension between assessing the old and the new is further complicated by state-to-state differences in content and instructional strategies.

NAEP Content Must Be Comprehensive

While there is no perfect solution to dealing with these tensions, NAGB, in its *Future of NAEP* report, has charted a wise course in its decision to align NAEP with national content standards while at the same time assessing the outcomes of current practice.[3] The Panel concurs with this policy but emphasizes that *the goal should be to ensure that the content of NAEP is comprehensive, rather than simply reflecting a balance or compromise between competing perspectives*. A comprehensive assessment, for example, should be able to detect an increase in computational skills at the expense of problem-solving abilities, or vice versa.

NAEP Frameworks Must Reflect National Performance and Content Standards

NAEP performance standards must be congruent with national performance standards and faithful to the substantive expectations underlying national content standards. One of the reasons for having national standards is to convey to the public the characteristics of good teaching and student learning. Verbal descriptions, sample tasks, and scoring criteria reported along with assessment results are the basis for communicating concretely what the standards mean.

The conclusion that NAEP frameworks will need to change to reflect emerging national content standards means that NAEP will have to stretch to encompass new substantive expectations. The practice of reporting assessment results using a single score needs to be reconsidered. *Changes in the mix of old and new content over time will almost certainly affect the interpretation of trends. One alternative is to report progress against the traditional and new contents separately*. This approach is consistent with NAEP's current practice of reporting both old and new trend lines. In the same vein, the Goals 3 and 4 Technical Advisory Group recommended that "data be partitioned so that the reader can discern how students are doing on items aligned with the NCTM Standards and those aligned with other standards."[4]

[3] National Assessment Governing Board, *The Future of NAEP* (Washington, D.C.: Author, August 1992), 10.

[4] Goals 3/4 Technical Advisory Group meeting, Council of Chief State School Officers (Washington, D.C., April 29, 1993).

NAEP has made substantial progress in developing forward-looking content frameworks for its assessments, aided by such external efforts as the NCTM Standards. Unfortunately, the press for frameworks to be developed and assessments planned and administered, often in a matter of 2 to 3 years, does not allow for adequate care to be taken at each stage. Externally developed content frameworks in some disciplines may be updated or revised faster than assessments can be planned. Thus the issue of continuity remains, both *within* an assessment, from framework development to the reporting of results, and *between* successive assessments in the same content area.

Standing Subject-Matter Panels Should Be Established

In its evaluation of the 1990 TSA, the Panel recommended that a mechanism be devised to ensure continuity, beginning with the development of a content framework and continuing through the various stages of item specifications, item writing, item scoring, and reporting of the results.[5] A set of oversight committees to achieve this objective was proposed as one such mechanism, but to date this recommendation has not been implemented.

This recommendation was elaborated upon in the Panel's *Setting Performance Standards for Student Achievement* and is reiterated here in slightly expanded form.[6] The Panel repeats the recommendation because it believes this is the mechanism by which the comprehensive content envisioned in *Goals 2000* can be linked to NAEP while at the same time reflecting current classroom practice. **The Panel recommends that NAGB establish standing subject-matter panels. In addition to subject-matter experts and researchers, membership in the panels should include classroom teachers, members of the public with expertise in the subject area, and representatives of NAGB. These standing panels should participate in the development of new frameworks and, when relevant, should consider explicitly how to incorporate certified national standards. Given the likely difference between the content based on nationally certified standards and that based on current classroom practice, the panels also should advise NCES on whether to combine assessment results from the two content domains into a single composite score or whether to report them as two separate scores. Finally, the panels should provide continuity to the assessment by being involved in all aspects of the process, including formulating the framework and objectives; reviewing items, item-scoring rubrics, and reporting formats; and helping to achieve agreement on narrative descriptions of performance standards and representative illustrative tasks. This would help to ensure that there is both logical and statistical correspondence between content standards, performance standards, and illustrative tasks.**

[5] The National Academy of Education, *Assessing Student Achievement in the States* (Stanford, CA: Author, 1992), 30.

[6] The National Academy of Education, *Setting Performance Standards for Student Achievement* (Stanford, CA: Author, 1993), xxix.

11 *Can NAEP Be Linked Meaningfully to State and International Assessments?*

The nation's concerns with how well its students measure up in comparison with students in other industrialized countries has fueled an interest in linking state and national assessments with international assessments. The states are keenly interested in linking their state assessments to international assessments such as the upcoming Third International Mathematics and Science Study (TIMSS) as well as to NAEP.[1] Meaningful links between state assessments, NAEP, and international assessments such as TIMSS would allow for the type of vertical integration of assessment results that NAEP alone cannot provide, in part because of the prohibition against reporting data at the school and district levels. Such links could make comparable performance data available at the international, national, state, district, and school levels. If linking could be shown to be technically feasible and valid, the benefits to states, districts, and schools would be immense, thereby supporting the Panel's policy relevance principle. As will be explicated in the following sections, the ability to make ***meaningful*** links between assessments presents considerable, although not insurmountable, challenges.

Equating Tests with Common Frameworks with Relative Ease

Tests and assessments designed to a common framework can be linked with relative ease. For example, it has long been common practice to use different forms of college and professional school admissions tests for reasons of test security, and the tests are then equated so that the results from the different versions are interchangeable. Similarly, major test publishers routinely publish alternative forms of their achievement tests, which are equated to a common scale.

The Challenge of Linking Assessments With Different Frameworks

In recent years, interest has increased in seeing whether useful approximations to equating can be realized when assessments are *not* designed to the same frameworks and when the content areas may be weighted rather differently between the

[1] When states not participating in the 1992 Trial State Assessment were asked what changes in NAEP might increase the likelihood of their future participation, over half indicated that the ability to link their state assessment to NAEP would be of importance.

assessments. As mentioned above, this interest in linking has been stimulated by the desire of states to measure their students' progress against emerging national performance standards and the desire at both the state and national levels to compare student performance to that of students in other countries.

The pending *Goals 2000* legislation would encourage states to develop their own content and performance standards as well as their own assessments for measuring progress against the goals. States or groups of states would be encouraged to develop assessments that would be linked to national content standards. Thus, a system of national assessments rather than a single assessment would be developed. For the different assessments to yield comparable results vis-a-vis the national standards, however, it must be possible to link all of them.

The Research Base

The Panel examined a study conducted by the Technical Review Panel (TRP), which monitors NAEP on behalf of NCES.[2] The study considered the feasibility of linking in situations in which all of the requirements of equating could *not* be met. The Panel also commissioned a companion study to better understand the influence of content on the dependability of such linkings.

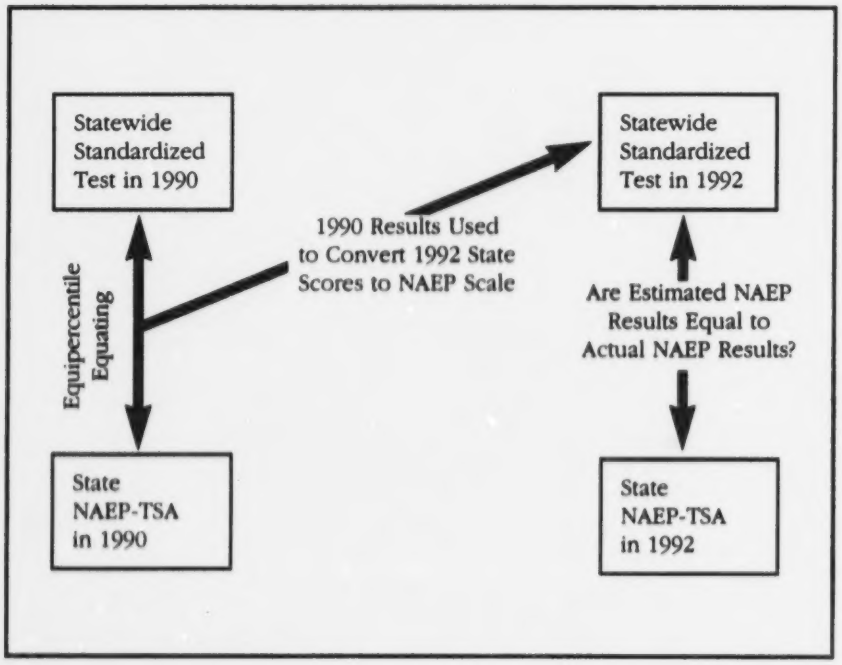
Linking Standardized Mathematics Tests to the 1990 NAEP Mathematics Assessment

In the TRP study, Linn and Kiplinger examined the linking of standardized mathematics tests to the 1990 NAEP mathematics assessment in four states. The data were obtained from four states that had participated in both the 1990 and 1992 TSAs and also had conducted statewide administrations of standardized achievement tests in each of those years.[3] For each state, the investigators equated the 1990 eighth-grade statewide achievement test results to the eighth-grade 1990 NAEP scale for mathematics and then used the data collected in the 1992 TSA to evaluate the accuracy of the 1990 equating. More specifically, as depicted in figure 11.1, an equating was performed to convert the 1990 standardized test results to the NAEP scale. That same conversion was then applied to the 1992 standardized test results to provide estimates of 1992 NAEP results, and those estimates were then compared to the actual 1992 NAEP results obtained in the TSA.

[2] R.L. Linn and V.L. Kiplinger, "Linking Statewide Tests to the National Assessment of Educational Progress: Stability of Results" (Boulder, CO: Center for Research on Evaluation, Standards, and Student Testing, 1993).

[3] Two states used the Stanford Achievement Test, one the Iowa Tests of Basic Skills, and one the California Achievement Test.

Figure 11.1. ***Schematic representation of study design***


Statewide Standardized Test in 1990
Statewide Standardized Test in 1992
1990 Results Used to Convert 1992 State Scores to NAEP Scale
Equipercentile Equating
Are Estimated NAEP Results Equal to Actual NAEP Results?
State NAEP-TSA in 1990
State NAEP-TSA in 1992

As Linn and Kiplinger state, "If conditions required for equating are completely satisfied, then any changes in the mathematics achievement of students within a state between 1990 and 1992 should have comparable effects on both NAEP and the statewide test results and, therefore, the equating obtained with 1990 data should still hold in 1992."[4] Unfortunately, in general the equating did not hold, and analyses showed differences larger than chance along most parts of the achievement distribution.

The pattern of results Linn and Kiplinger observed was also of special interest to the Panel because the 1992 NAEP scores were lower than would be expected in the state where the standardized test was used for the first time in 1990 and the third time in 1992, but were higher than expected in the state where the standardized test was used for the first time in 1992. That is, students "did better" relative to NAEP in those states where the standardized test had been used for 1 to 3 years compared to the results where the standardized test was used for the first time. These results suggest that teaching to the test had occurred, but the "gains" on the standardized tests due to teaching to it did not generalize to other tests, NAEP in this case.

[4] Linn and Kiplinger, op. cit., 16.

Examining the Content Congruence Between the Standardized Mathematics Tests and the 1990 NAEP Mathematics Assessment

The companion study commissioned by the Panel examined the content congruence between two of the standardized mathematics tests used in the Linn and Kiplinger study and the 1990 NAEP mathematics assessment item pool.[5] In this study, a panel of content experts in mathematics classified the items from each of the standardized tests into NAEP's five mathematics content categories and three ability categories.[6] The results showed that disproportionately large numbers of items were classified into the combined numbers and operations/procedural knowledge and numbers and operations/conceptual understanding categories. This means that the standardized tests did not assess processes such as problem solving, or content such as geometry or probability and statistics, to the extent that they are covered on NAEP.

The differences in content coverage between the standardized tests and NAEP provide a likely explanation for the finding of Linn and Kiplinger that trustworthy equatings of standardized tests to the 1990 and 1992 NAEP mathematics assessments could not be achieved. Taken together, the empirical results from these two papers support the theoretical work of Mislevy and Linn,[7,8] which states that, of the various types of linking, only equating allows one to use two forms of a test or assessment with indifference.

Taken as a whole, the results from the Panel's study and the Linn and Kiplinger study suggest that *state assessments using standardized achievement tests may not be able to establish a trustworthy link between their assessments and NAEP. Furthermore, by implication, states will be better able to equate their state assessment results to NAEP if their assessments weigh content areas in a manner consistent with the NAEP frameworks.*

The Panel's results should not be surprising. It is well known, for example, that when major changes are made to the framework test or assessment, it is wise to construct a new scale. It is for this very reason that NAEP introduced new scales for mathematics in 1990 and reading in 1992. If new scales were not constructed and new trend lines begun when major changes were made to the framework, comparisons based on trend data would be quite meaningless.

[5] L. Bond and R.M. Jaeger, "The Judged Congruence Between Various State Assessment Tests in Mathematics and the 1990 National Assessment of Educational Progress," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

[6] The five content categories are numbers and operations; measurement; geometry; data analysis, statistics, and probability; and algebra and functions. The three ability categories are conceptual understanding, procedural knowledge, and problem solving.

[7] P.J. Mislevy, "Linking Educational Assessments: Concepts, Issues, Methods, and Prospects" (Princeton, NJ: Educational Testing Service, December 1992).

[8] R.L. Linn, "Linking Results of Distinct Assessments," *Applied Measurement in Education* (6) (1993): 83-102.

Requirement of Common Frameworks for Linking

While more research is needed, *it seems likely that a common framework, along with items that are sampled from that framework, are necessary conditions for linking state assessments to NAEP.* However, because the reliabilities of two assessments, their distributions of item difficulties, and the context within which they are administered (e.g., the stakes) can all influence the ability to equate assessments, it seems *un*likely that the use of NAEP frameworks will be the single *sufficient* condition for adequate equating results. An ongoing study to equate the 1992 Kentucky state assessment to NAEP should shed additional light on this issue, however.

The Prospect of Linking NAEP to International Assessments

Linking NAEP to international assessments (e.g., TIMSS) raises similar technical issues and a major policy question as well. As discussed above, the issues of equivalence of frameworks, content, and reliabilities, plus sampling and item-scoring complexities, present major technical problems to be overcome. Although it may be difficult to link NAEP and TIMSS in a trustworthy manner, such a linking is highly desirable because it would greatly enhance the value of both the national and state assessments. Hence, it is critical that the basis for a sound linking be built into the plans for the TIMSS. **Based on its research, the Panel recommends that TIMSS either use a framework that is similar to the NAEP framework, or that a subset of the TIMSS assessment be mapped onto the NAEP framework so that scores with comparable weighting of content from the TIMSS can be equated to NAEP.**

Finally, while a common framework and items chosen in roughly the same proportion would seem to be necessary conditions for linking NAEP or state assessments to TIMSS, the meaning of observed performance differences between the United States or individual states and other countries will also depend upon understanding the considerable differences among nations, including differences in curriculum, scope and sequence of instruction, distribution of opportunities to learn, instructional methods, rewards for superior performance, and cultural attitudes toward education.

In summary, the desirable objective of linking NAEP and state assessment data to international assessments such as the TIMSS presents major challenges. However, with proper planning, careful attention to the content coverage of the two assessments, and coordination, trustworthy links should be attainable.

12 How Can State Participation in State NAEP Be Maximized?

As discussed in chapters 4 and 6, many states strongly support state NAEP and find it useful for monitoring student performance and fostering education reform. Nevertheless, the 1992 state NAEP made it clear that a number of obstacles to the participation of states and individual schools remain. Only 37 states and jurisdictions participated in state NAEP in 1990, and while the number increased to 41 in 1992, this is still substantially fewer than all 50 states. Furthermore, the proportion of initially sampled schools refusing to participate in state NAEP actually increased from 1990 to 1992, on average,[1] and refusals rose to unacceptable levels in several states.

The Burdens States Face in Participating in State NAEP

One of the most difficult tasks the states face in participating in state NAEP is meeting the contractor's requirements for district and school participation. The results of the assessment are credible to the degree that these sampling requirements are met; however, unless the state legislates participation in NAEP (as some do), there is little direct incentive for the districts and schools to participate.

The burden of securing participation is doubly difficult for small states. Not only must the same schools be recruited year after year because there are not sufficient schools to serve as alternates, but when multiple subjects are assessed at the state level, some 4th, 8th, and 12th graders may have to be assessed in more than a single subject. Multiple subjects are likely to become the norm once state NAEP is well established.

In authorizing state NAEP for 1994, Congress followed the Panel's advice that private school students be added to the assessment. The addition of private schools to the sampling frame will undoubtedly make the task of securing cooperation in 1994 greater than in 1990 and 1992. In the case of public schools, a state can use a combination of positive and negative sanctions to increase participation; but because of legal requirements, many states cannot use state funds for recruiting private schools. Moreover, certain classes of private schools are unlikely to be willing to participate due to their views on large-scale testing.

To date, the costs for the development, scoring, analysis, and reporting of state NAEP have been borne by the federal government, with cost of administration shared between the federal government and the states. The costs the states have been required to bear have been primarily in-kind costs, although these can be considerable, as the Panel's cost study has shown. The U.S. Department of Education's draft plans for the reauthorization of NAEP give the Secretary of Education

[1] As discussed in chapter 6, however, the high initial school participation rate (94 percent) of 1990 may have been unusual and related to the novelty of the undertaking.

the option of shifting some or all of the costs for scoring, analysis, and reporting for state NAEP to the participating states. It seems likely that disincentives to the states to participate in state NAEP will increase proportionately to the degree that the Secretary uses this option.

Incentives That Might Be Employed to Maximize Participation in State NAEP

The Panel has considered a variety of incentives that might be used to maximize participation in state NAEP. Those that appear to offer the best hope for success, or that have been most widely discussed in the NAEP community, are presented here.

Remove the Prohibition Against Reporting Data Below the State Level?

Given the problems in securing district and school participation noted above, one obvious incentive to maximize participation would seem to be to remove the prohibition against reporting data at the district and school levels. As pointed out in chapter 9, however, compelling reasons exist for not lifting the prohibition. Moreover, allowing districts and schools to report NAEP results would likely provide incentives for only those very few large districts and schools that could afford to pay the cost, while reporting results could actually be a disincentive for schools that perform poorly. It is plausible, therefore, that removing the prohibition would have little effect on the state coordinators' abilities to recruit average schools and school districts to participate in state NAEP. Even if it could be demonstrated that removing the prohibition against district- and school-level reporting *did* increase the motivation for participation, the gains in participation would have to be weighed against the possible corruption of NAEP that might occur as a function of removing the prohibition. The Panel believes that the overriding consideration must be the protection of the integrity of national NAEP. In sum, ***the Panel does not believe that if the prohibition were removed, the increase in districts' and schools' motivation to participate would outweigh the possible damage to NAEP.***

Allow States to Mix and Match Subjects and Grade Levels Assessed? Move to Annual NAEP Assessments?

The CCSSO Education Information Advising Committee (EIAC) subcommittee on assessment has recommended that states be allowed to choose the subjects and grade levels to be assessed as a way to reduce the states' burden in participating in state NAEP.[2] The subcommittee also recommended that NAEP

[2] Passed as part of a resolution on the future of NAEP at the EIAC meeting on September 17, 1992.

legislation be amended to allow for annual assessments for those states that volunteer to participate.

The Panel has considered both of these recommendations and believes that the option of "mixing and matching" subjects and grades has merit but also has disadvantages that outweigh the merits. In spite of whatever limitations state-by-state comparisons might have (and as the Panel pointed out in its evaluation of the 1990 TSA, there are serious limitations[3]), part of the public appeal for state NAEP is based on the reporting of data for *all* participating states. If states were allowed to pick and choose the subjects and grades in which they participate, state-by-state reporting could be rendered totally meaningless. And, without the publication of state-by-state results, public support for state NAEP could be severely eroded. As an example, consider the results by ethnic groups for the 1992 reading assessment. In these comparisons, only students of similar racial-ethnic background are compared, and the results showed very different patterns of performance in different states. For example, in California, the analyses showed that fourth graders in every racial-ethnic group scored very poorly, placing them in the lowest fifth (quintile) of participating states. By contrast, white students in Washington, D.C. performed very well compared to white students in other states (top quintile), while D.C. minority students, like those in California, scored in the lowest quintile.[4] Whatever limitations may be attached to it, a quintile comparison of this sort makes sense only if the vast majority of states is required to participate in all subjects at all grade levels when they sign up to participate in state NAEP.

CCSSO and EIAC have also recommended that NAEP assessments be conducted annually instead of biennially, with the intention that fewer subjects be assessed in any given year. **While it is not in the Panel's purview to decide on which subjects should be assessed and when, the Panel recommends allowing NAEP assessments to be administered each year as a way to effectively distribute the administrative work over the years and reduce the burden on states in any given year.**[5]

Allow Reduced Student Samples for Small States?

One proposal being considered by NAEP would allow small states to use smaller samples (e.g., half the number) of students in sampled schools.[6] Analyses by Westat indicate that the use of even half samples in these states would not introduce an unacceptable level of sampling variation in published statistics based on NAEP. Contributing factors include the relative homogeneity of small states and the relative

[3] The National Academy of Education, *Assessing Student Achievement in the States* (Stanford, CA: Author, 1992), 47-50.

[4] I.V.S. Mullis, J.R. Campbell, and A.E. Farstrup, *NAEP 1992 Reading Report Card for the Nation and the States* (Washington, D.C.: National Center for Education Statistics, September 1993), 29.

[5] Both the Department of Education and the National Assessment Governing Board support moving to annual assessments as well.

[6] Small states are defined as states where the sample chosen for a given grade is as large as the population of students at that grade level.

size of the sample compared to the whole population of the state. Similar results would not necessarily follow if samples were reduced in larger states. **Assuming that stable statistical estimates can be based on samples of reduced size, the Panel recommends that small states be allowed to use reduced samples of students when participating in state NAEP.**

Use Different Assessment Administrators?

Under current arrangements, participating schools are required to provide assessment administrators, unlike national NAEP in which the contractor provides the administrators. However, current rules do not allow the teacher of the students being assessed to be their assessment administrator. In response to a questionnaire sent by the contractor seeking information about how to increase district and school participation rates, the state coordinators mentioned, among other suggestions, replacing the current assessment administrators either with NAEP staff or with the students' teachers.[7] With regard to the latter suggestion, the Panel concludes that using the students' teachers is not a good idea. While the vast majority of teachers would be impeccable in administering the assessment, a few would be tempted to engage in behavior that might improve their students' scores.

Cost was the main reason Congress mandated the use of local assessment administrators rather than NAEP personnel. While the Panel believes that using NAEP personnel might increase participation rates marginally, given the findings in chapter 2 that the scores of the monitored and unmonitored sessions are virtually identical and that the average NAEP scores of the TSA and the composite national sample are very similar, the increased cost of using contractor personnel to conduct the TSA administrations would seem to be unwarranted. Therefore, the Panel suggests continuing the use of local assessment administrators.

Have the Federal Government Continue to Bear a Major Portion of the States' Costs for the Assessment?

The U.S. Department of Education's draft legislation would give the Secretary of Education the authority to require that states pay for some or all of the costs of scoring, analyzing, and reporting state NAEP data—costs currently paid by the federal government. The proposed legislation has both advantages and disadvantages. On the positive side, should the budget for NAEP ever become so small that it would be otherwise impossible to implement state NAEP, the cost sharing might be embraced by some states. However, a number of problems could also arise. First, if states do decide to pay a larger share of the costs, they would be in a stronger position to argue that they should be able to choose the subjects and grades assessed. The drawbacks of this so-called "mix-and-match" plan were discussed in a preceding section. Second, with a larger cost burden, some, if not many, of the states that might otherwise have

[7] National Center for Education Statistics, *1992 TSA: Summary of Participation Rates, Training, and Data Collection Activities* (Washington, D.C.: Author, July 1992), 20.

participated, would not. (As evidence for this concern, the Panel notes that when states that did not participate in the 1992 TSA were asked what incentives might be used to increase the likelihood of future participation, two-thirds of the responding states mentioned *lowering* their costs through reimbursement for part of the cost for administration and paperwork.[8]) Of course, the smaller the number of states that participate in state NAEP, the more difficult meaningful cross-state comparisons become. Third, once the option of moving some of the cost to the states is exercised, it may become difficult to move those costs back to the federal government, thereby creating a permanent disincentive for state participation in state NAEP.

The Panel recommends that the federal government not move additional costs for state NAEP to the states. While the motivation for the proposed legislation may be to protect state NAEP, implementation of the option to bill the states for NAEP-related costs might actually result in fewer states participating in state NAEP.

[8] E. Hartka and D.H. McLaughlin, "A Study of the Administration of the 1992 National Assessment of Educational Progress Trial State Assessment," in *The Trial State Assessment: Prospects and Realities: Background Studies* (Stanford, CA: The National Academy of Education, forthcoming).

13 Prospects and Realities

In this chapter, the Panel draws on the guiding principles presented in chapter 5 to consider a design for a comprehensive NAEP—a NAEP unencumbered by current fiscal restraints. However, because budgetary considerations are a reality, the Panel also examines how these guiding principles can best be realized in light of restrictive federal budgets. Building an effective NAEP in a time of constrained resources requires close examination of potential cost savings and trade-offs among various objectives for NAEP.

A Design for an Effective and Comprehensive NAEP

In this section, a design for NAEP is presented based on the assumption that the budget for NAEP would be less constrained than it is at present. The purpose in presenting this design is to provide Congress with the Panel's view of what would comprise a comprehensive NAEP if increased funding is available in future years. The issues considered include how often NAEP would be administered, who should pay for the cost of participating states, who should be assessed, and how the results should be reported.

NAEP Would Be an Annual Assessment with Comprehensive Content and Frameworks

In order to assess the full range of subjects envisioned by NAGB, national and state assessments would be conducted annually in one or more subjects at grades 4, 8, and 12. Administering selected assessments every year would reduce the testing burden in any given year, something particularly important for state NAEP where the available sampling units are more limited. Following the Panel's lead and reflect principle, the frameworks and their content would be comprehensive; that is, the frameworks would reflect up-to-date theory and research in NAEP content areas while at the same time mirroring the best of current classroom practice. The frameworks would also incorporate evolving national content and performance standards as they are developed and certified. Both classroom practice and content experts' views of what should be taught can be expected to change over time; however, both types of changes are likely to evolve slowly. This gradual evolution is important because the establishment of meaningful trend lines requires that frameworks remain relatively unchanged for an 8- to 10-year period.

NAEP's Role as an Independent Monitor Would Be Protected

The unique role of NAEP as the nation's best independent monitor of the achievement of the nation's students and of progress toward meeting national goals needs to be

protected. An independent and uncorrupted monitor is essential given the current push to expand the role of assessment for ensuring school accountability and for determining whether individual students are meeting high levels of performance on challenging material determined by national content standards. Past experience in this country as well as other countries has repeatedly demonstrated that the same assessment cannot simultaneously be an instrument of accountability for schools, teachers, or individual pupils and an independent monitor. Without a dependable, independent monitor, it is impossible to determine the degree to which changes observed on instruments of accountability are valid reflections of changes in student achievement or are artifacts of narrow attention to specific assessment tasks (and formats). Hence, it is critical that NAEP's role as an independent monitor be protected so that there can be clear evidence of the nation's progress in reaching achievement levels related to its educational goals.

The Cost for State Participation Would Be Borne by the Federal Government

Participation in state NAEP would be encouraged by having the full cost of the assessment paid by the federal government. In exchange, states participating in a given year would be required to participate in all subject areas assessed, and at all grades assessed, thereby increasing the utility of the results for the nation and the states.

The Assessment Would Be More Inclusive

The purpose of NAEP is to provide information about what the nation's 4th, 8th, and 12th graders know and can do; therefore, the greater the scope of sampled students actually assessed, the better the estimate of student achievement at those grade levels. In a comprehensive NAEP, private school students would continue to be included at both the national and state levels, as they are in 1994. Children with special needs (IEP, LEP) would also be included, to the extent feasible. Studies would be done to determine the feasibility of administering special versions of NAEP to students with disabilities, including individually administered, Braille, large-type, and computer-assisted versions of the assessment. As more information is obtained about the IEP and LEP population presently excluded from the assessment, new, more inclusive guidelines would be set, commensurate with ethical, technical, and financial considerations. Foreign-language versions of NAEP would also be available on demand for use by the many students whose native language is not English. Initially, a Spanish-language version might be available, with other language versions being added as feasible.

Finally, while it may be prohibitively expensive for NAEP to assess out-of-school children, special samples of unenrolled, nongraduated 17- and 18-year-olds would be assessed with blocks of NAEP items, perhaps as part of a periodic national adult literacy study. If the population of out-of-school eighth graders continues to grow, a special sample of this age group would also be drawn for assessment and study.

The Assessment Would Report What Students Should Know as Well as What They Actually Know

Results from national and state NAEP would be reported both in terms of what the nation's and participating states' children actually know and what they should know. That is, not only would average scores and anchor descriptions of performance be provided, but results would also be reported in terms of progress against challenging standards and high levels of achievement. Both national and state NAEP would be linked to available major international assessments to increase our understanding of how U.S. students compare with students in other countries. This information would also be used to determine the levels of attainment of the best students in nations around the world.

NAEP Would Continue to Be a Leader in Assessment Innovation

NAEP has historically been a leader in assessment innovation; however, the challenges of assessment continue to grow with the demand for better measures of integrated problem-solving and other higher order cognitive skills. Especially important is the need to be able to measure problem-solving skills at very advanced levels. NCES should continue to explore the psychometric and design issues surrounding the increasing use of performance-based and individually administered exercises to measure these skills at all levels of difficulty. Other areas for research include how to scale items and content areas (e.g., mathematics) that appear to be inherently multidimensional and how to design assessments for children with disabilities, as well as assessments in languages other than English.

Difficult Choices Given Current Funding Realities

The Panel is aware that the full scope of NAEP described in the preceding section requires substantial fiscal resources. For this reason, in this section the Panel considers whether significant cost savings might be realized by eliminating the current national NAEP sampling frame and replacing it with a sampling frame that generates national NAEP performance estimates from state NAEP samples. After considering this issue, the Panel briefly discusses the difficult choices that NAEP faces and concludes with a recommendation for NAEP funding.

The Possibility of Aggregating State NAEP Samples to Obtain National Performance Estimates

It has been suggested both to and within the Panel that mounting two assessments—national and state NAEP—is not cost efficient. A seemingly obvious solution would be to aggregate the results from state NAEP in such a way as to create national NAEP estimates. Supplemental samples would probably have to be drawn from those states not participating in state NAEP.

The first step in deciding whether this suggestion is feasible rests on examination of the purposes and uses of national and state-level data. As has been stated in both chapters 1 and 5 of this report, NAEP is the single best indicator of what the nation's children know and can do. Therefore, it is imperative that the integrity of national NAEP be preserved. There may be content areas of interest to Congress that may not be of high priority for the states. Furthermore, even within a given subject area, national NAEP is not merely a single sample that parallels that of the TSA. National NAEP includes a large number of bridge studies used for charting trends, as well as other special studies conducted on a periodic or as-needed basis. It would not be cost effective or, in some cases, even feasible to construct the samples for these ongoing and special studies by aggregating the state NAEP samples.

For all these reasons, it seems unlikely that national NAEP could or should be totally replaced. However, cost efficiencies might be gained using an aggregation of state NAEP samples to obtain national population estimates in *selected grades and content areas*. In those grades and content areas where the federal government and the states are in alignment with respect to areas to be assessed, consideration could be given to substituting an aggregation of state samples for the current national NAEP sampling framework.

Aggregating state NAEP samples to obtain national estimates has other problems, however. First, the level of voluntary participation by states year after year in assessments involving multiple content areas and grades remains unknown. The fewer the states that participate in state NAEP, the more inefficient the idea of deriving national performance estimates from an aggregation of state NAEP samples becomes.

Second, states would have to make firm commitments to participate in NAEP several months in advance. A backup process would need to be in place in case states decided to drop out at a late date or in case they failed to achieve minimum school participation requirements.

Third, the cost of state NAEP is very high relative to the cost of national NAEP. For 1994, the cost of state NAEP for one subject at one grade level is estimated to be $9-10 million.[1] Adding either another subject at the same grade level or a second grade level adds an additional $4-5 million to the cost of the assessment. The cost of national NAEP for three grades and three subjects for 1994 is estimated at $10-12 million. The cost of aggregating the 1994 state NAEP into a national NAEP (assuming

[1] The cost information in this section comes from discussion with NCES staff. The cost figures are for data collection and analysis only. The development costs, which run from $3-4 million per subject, exist whether there is a state NAEP or not.

three subjects were assessed at three grade levels) would be between $41 and $50 million, which *excludes* the cost of drawing supplemental samples for nonparticipating states.[2] Hence, the cost of a national NAEP aggregated from a state NAEP consisting of three subjects at three grade levels would be more than four times the cost of the current national NAEP and would be reasonable only if state NAEP were fully funded.

In summary, it seems unlikely that national NAEP could ever be fully replaced by aggregating state samples. However, if state NAEP were conducted on an annual basis with a high level of state participation, cost savings *might* be realized and the burden on states reduced by estimating performance on national NAEP through an aggregation of state NAEP samples. Therefore, **the Panel recommends that NCES examine whether cost efficiencies could be realized by a partial substitution in which aggregation of state NAEP samples would be used to make national estimates of student performance, supplemented as necessary with samples from nonparticipating states. The aggregation of state results for national estimates in a given grade and content area would not rule out the possibility of national NAEP being carried out in other content areas or at other grade levels.**

Maintaining the Quality of Both National and State NAEP

Any trade-offs regarding cost must be made not only in the context of maintaining the present quality of NAEP, but also with the view to enhancing it. As discussed throughout this report, the challenges to NAEP's technical capacities continue to grow. NAEP must, for example, update and expand its content coverage to encompass the new national content standards currently under development. Appropriate assessment of higher order cognitive skills and the upper reaches of achievement, in line with emerging national performance standards, will also require new item development as well as possible special testing of the nation's most able students.

In addition, NAEP's paramount monitoring function, as well as the inclusiveness called for in the Administration's *Goals 2000* legislation, argue for expanded coverage of special, hard-to-assess populations. Throughout this process of growth and development in the nation's educational aspirations, NAEP has the opportunity to serve as a leader in assessment innovation. Such innovation implies greater development costs and, perhaps, greater administration costs as well (e.g., for scoring performance tasks).

The value of all these activities must be weighed against the desire for assessing more subjects at the state level. The substantial value participating states see in state NAEP was an important factor in the Panel's decision, discussed in chapter 6, to endorse the authorization for a continuing state NAEP. However, the Panel's recommendation rests on the assumption that adequate funding is available for *both* state and national NAEP. The Panel is not in favor of constraining the quality of what is currently measured by the national assessment in order to assess more subject areas at the state level, for example.

[2] This figure is derived by assuming that the first subject at one grade level costs $9-10 million and adding to it $4-5 million for *each* of the additional eight subject-by-grade assessments.

More funding is needed if the technical quality of both state and national NAEP is to be maintained. The 1994 NAEP had been planned for five subjects at the national level and two at the state level—a reasonable plan in the Panel's view. However, the funding appropriated for 1994 NAEP will allow for only three subjects at the national level and one at the state level, and in the latter case, at only one grade level. Therefore, the full potential for both national and state NAEP is far from being realized. **The Panel recommends that Congress provide adequate funding for future NAEP assessments—funding that will allow national NAEP to continue to be a high quality indicator of the educational achievement of all the nation's children while at the same time providing for an annual state NAEP in *at least* one subject at three grade levels. The funding required to achieve these goals is relatively modest when compared to the benefit that the nation and states will realize from the assessments.**

A Concluding Thought

The Panel has witnessed a significant change in attitudes toward the value of state NAEP over the past 3 years. While several states were supportive of a state-by-state NAEP when it was put into law in 1988, many others were skeptical of its value. Today, the vast majority of states see value in state NAEP; as a result, it is possible to envision the day when nearly all, if not all, states will participate.

Part of the growth in the perceived value of NAEP for the states is undoubtedly related to the appearance of the TSA at a time when the education reform movement began to gain substantial momentum. For many states, the TSA has provided a baseline showing the relatively poor mathematics and reading performance of their students. State NAEP, if continued, will provide these same states with a state-of-the-art assessment tool for tracking the educational progress of their students across time. Moreover, state NAEP can provide the states and the nation with information about participating states' progress against national education Goals 3 and 4. Finally, assuming international assessments such as the upcoming Third International Mathematics and Science Study are designed in ways that allow linking with state NAEP, states will be able to chart their progress against international standards as well.

Even with these attributes, as this evaluation has documented, bothersome technical problems with state NAEP remain that must be resolved before this experiment can be judged a total success. Furthermore, it has yet to be demonstrated that states can carry the burden of assessing two or three subjects at three grade levels. These concerns notwithstanding, the Panel is enthusiastic about the potential for state NAEP for tracking the educational progress of the states, not only against their own education standards, but also against those of the nation and the world. Based on the Panel's findings, state NAEP should be continued. However, ongoing evaluation is also necessary with respect to the issues raised in this report.

Finally, it should be emphasized that state NAEP is no substitute for assessments used by teachers in the classroom and by district and state agencies. Education reform is

being accelerated by informed and innovative school-level activity that integrates assessment procedures with teaching practices. NAEP should not be part of these developments; rather, it should serve as a monitor that assesses the *results* of education reform and that provides information to the education community about its successes and about areas in need of further effort.

Appendix

Synopses of Studies for the National Academy of Education Panel on the Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment

This appendix contains summaries of 11 studies that were commissioned by the Panel and served as the basis for the Panel's findings and recommendations. The complete reports are being published in a separate volume, ***The Trial State Assessment: Prospects and Realities: Background Studies***.

A Study of Eligibility Exclusions and Sampling: 1992 Trial State Assessment

Bruce D. Spencer
Northwestern University

The 1992 Trial State Assessment (TSA) of the National Assessment of Educational Progress (NAEP) tested samples of the fourth- and eighth-grade public school population in 41 states, Washington, D.C., and 2 U.S. territories. The assessment excluded certain large groups of students, some by design (e.g., nonpublic school students and students with limited English proficiency [LEP] or individualized education plans [IEP]) and some by chance (nonparticipants, movers who "fell through the cracks" of the survey procedure, and absentees). In addition, the TSA excluded former students who dropped out of school. This study examines issues related to the design of the sample survey, unintended exclusions of absentees and other students, and deliberate exclusions of students.

Deliberate Exclusions

The TSA was designed to demonstrate the feasibility of conducting a statewide assessment in a consistent and controlled manner across states. The TSA covered such a large number of states and schools within states that the number and diversity of jurisdictions covered is more than adequate for the demonstration purposes. However, the design of the TSA allowed for the exclusion of three groups of students:

- students enrolled in private schools
- students with limited English proficiency who were judged to be incapable of participating meaningfully in the assessment
- students with individualized education plans who were judged to be incapable of participating meaningfully in the assessment.

Nationwide, one in five or six students belongs to at least one of these three groups. Differences among the states in the proportions of students in these groups could have important effects on state-by-state and state-to-national NAEP comparisons. In addition, users could draw misleading inferences from the TSA data by believing, for example, that the percentage of *all* public school students performing at a given ability threshold is higher than it actually is when students with IEPs or LEP are included.

Private school students are estimated to make up about 11 percent of the fourth-grade and 10 percent of the eighth-grade populations nationally; however, the percentages vary widely across states. Because private school students were not tested in the TSA, illustrative sensitivity analyses were conducted to suggest how large the effects of their exclusion might be, assuming that the differences between public and private school

students were constant for all states within each of four regions (northeast, southeast, central, and west). The results indicate that the changes in rankings would mostly be modest, with an average change in rank of less than 1. The effects on the estimated proportions of students in a state at various mathematics achievement levels are small, for both grades 4 and 8.

The exclusion criteria for LEP and IEP students were part of the sampling design and were implemented locally. Students with disabilities who are not enrolled in regular public schools were also excluded. Overall, about 2-3 percent of the students in the fourth- and eighth-grade samples were excluded because they had limited English proficiency, and about 4-5 percent were excluded because they had individualized education plans. Percentages across states and territories varied: LEP exclusions ranged from 0-2 percent in most states but were 5 percent for grade 8 and 10-11 percent for grade 4 in California. IEP exclusions ranged from about 2 percent in North Dakota and the Virgin Islands to 7 percent in several states and the District of Columbia.

The exclusion rates from national NAEP and from the TSA seem consistent (within 1-2 percent of each other). TSA procedures require that any sampled students who are classified as IEP or LEP in their school records be identified as such on the assessment administration schedule, whether or not they are excluded.

Comparisons of the classification rates from the 1990 and 1992 assessments indicated that the rates were not highly stable (e.g., for grade-8 IEP, the correlation between 1992 and 1990 classification rates was .76). Further, compared to national NAEP, the TSA seems to classify too many students as IEP or LEP. In the TSA, about 14-15 percent of the fourth-grade students and about 13 percent of the eighth-grade students were classified as LEP or IEP; in national NAEP, about 12 percent were classified as LEP or IEP. For the 1990 grade-8 TSA, the discrepancy was even larger. The discrepancies indicate that the NAEP/TSA classification of students as IEP or LEP is not stable. As an alternative IEP classification rate, U.S. Department of Education statistics on the number of IEP students in each state by age were used in making comparisons, and the TSA IEP rates were found to be slightly lower. The variability of the rates reflects the well-known district-to-district variability in policies for classifying students as IEP more than just differences in student characteristics.

In investigations of the combined effects of private, IEP, and LEP exclusions, the effects of including all IEP students were found to be moderate. The rankings of the states would change on average by 1 or 2 places, with a few changing by as much as 4, 5, or 6. The effects of including LEP students are typically smaller, except for grade-4 scores in California, which would drop by 3-5 percentage points. Including both IEP and LEP students had a greater effect, causing drops of up to 6 points. Including IEP, LEP, and private school students changes rankings substantially, but typically by somewhat less than if just IEP and LEP students are included.

By design, NAEP and the TSA do not test young people who have dropped out of school, and dropouts tend to perform at lower scholastic levels than students who remain in school. Hence the expected TSA scores are higher than what they would be if dropouts were tested.

Sampling of Schools and Students

The school sample design was evaluated and found to be the standard kind of sample design used in high quality, major national education surveys. The investigator notes that it seems to have been executed with care and due attention to detail.

The NAEP contractor implemented supplementary sampling procedures to account for students who moved between the time the school lists were submitted and the time of the assessment, but these procedures seem to have accounted for only 60-70 percent of the students they were supposed to cover. The underrepresentation of transfer students may cause biases in the state scores. The investigator compared the scores of "mobile" students (those who changed schools at least twice in the past 2 years) with the scores of nonmobile students, and large differences were found.

Among selected students, participation rates were good, with absentees comprising the major portion of the nonparticipating students in the sampled (and participating) schools. Each school with at least three nonexcluded, nontested students was instructed to offer a makeup session, but relatively few of the initially absent students attended. Those who did attend scored lower than those assessed who had not been absent.

Weighting

Although the weighting to reflect unequal selection probabilities was reasonably straightforward, the weighting to adjust for student nonparticipation was crude because the characteristics of selected but nonparticipating students were not recorded. There is no reason to believe that the adjustments were highly effective in removing nonresponse bias.

Summary and Recommendations

On the basis of his analyses, the investigator made the following recommendations:

- Because there is no apparent justification for excluding private school students, they should be included in future assessments.
- If, as seems likely, the decision to exclude some IEP and LEP students is based on difficulty of assessing them, then the cost-benefit analysis implicit in the exclusion rules decision needs to be examined.

- If the exclusion rules are relaxed so that greater proportions of IEP and/or LEP students are to be tested, then modifications of the data-collection procedures, sampling design, and analytic strategies (including so-called small-area modeling techniques) may be appropriate.

- If the current inclusion and exclusion policies are continued, the reporting of NAEP results should be changed to make clear the implications of the exclusions. In addition, more study should be given to the formulation of consistent and valid criteria for exclusion and to the implementation of the criteria.

- Careful consideration should also be given to the assessment of dropouts. At a minimum, if dropouts are not to be included in the assessment, then the proportions of the age cohort who are dropouts should be provided to users of the data.

It is possible that the sample design could be made more efficient through improved stratification of schools.

- In future assessments, attention should be given to stratifying schools within a state by test scores on statewide exams.

- Transfer students seem to be markedly underrepresented in the assessment. More attention should be given to capturing transfer students in the assessment or, failing that, to making appropriate analytic adjustments.

Participation rates were reasonable for an assessment of this nature.

- In future assessments, some attention should be given to estimating the amount of systematic error caused by nonparticipation by schools and students within schools. Special attention should also be given to developing additional sources of data to adjust for school and student nonparticipation.

A Study of Students Excluded from the 1992 National Assessment of Educational Progress Trial State Assessment

Elizabeth Hartka
American Institutes for Research

The National Assessment of Educational Progress (NAEP) purports to be the nation's report card, reporting on the achievement of a sample representing *all* U.S. students. The sampling design provides guidelines for excluding certain students: those with individualized education plans (IEP) and those with limited English proficiency (LEP)

who are judged to be incapable of participating meaningfully in the assessment. Students who have dropped out of school are also excluded; private school students are excluded from the Trial State Assessment (TSA) but not the national assessment. This study focuses on the procedures for excluding IEP and LEP students from the TSA and on this question: Would the same IEP or LEP student have the same chance of participating in the TSA regardless of that student's state of residence? Given the paucity of state-by-state data available from non-TSA state assessments, the findings of this study can be illuminating, providing state-by-state analyses of exclusions made under comparable conditions using standardized training and assessment conditions.

Methods

Variables for analyses included performance data from the 1990 and 1992 TSAs in mathematics, selected student background variables, and data from the excluded student questionnaire, which should have been filled out for each student excluded from the TSA. This questionnaire collects information on the excluded student's gender; race or ethnic group; reason for exclusion (i.e., IEP or LEP); type and seriousness of disability if IEP; first language if LEP; percentage of time the student is mainstreamed in a regular classroom; and percentage of time the student spends in a special education program or in a special language program.

Two key dependent variables were used: (1) the percentage or number of IEP or LEP students excluded from the assessment in 1992, which was examined state by state, within and across TSA sessions in most analyses; and (2) the student performance score on the assessment, both overall and item by item.

A preliminary analysis focused on the consistency of proportions of excluded students between states participating in both the 1990 and 1992 NAEP assessments, with the rationale that abrupt changes in proportions might indicate changes in the way the exclusion rules were applied. Additional analyses looked at the relationship between session scores, school characteristics, and the proportion of excluded students. Comparisons of score distributions for IEP-LEP and non-IEP-LEP students were made to determine whether the TSA may have been too easy or too difficult for IEP-LEP students; similarly, scores for both groups on open-ended and extended-response items were compared. Finally, the data were analyzed for evidence of categorical exclusions by schools in which all *identified* IEP and LEP students were excluded from the assessment.

Results and Discussion

Due to the demographic correlates of limited English proficiency, one would expect states to vary widely in the size of their respective LEP populations, and this was the case. California, Florida, and Texas had the largest numbers of LEP students. The size

of special education populations also varied, ranging from 5.9 percent of sampled students in Hawaii to 16.3 percent in Massachusetts, with a majority of states reporting that more than 10 percent of their students are in special education classes.

Thirty-seven states and territories participated in both the 1990 and 1992 TSAs. Analyses of the changes in rates of exclusion indicated that most of the states' ranks were stable between the two assessments. Four states and territories changed their rank orders quite dramatically; however, the changes in the percentage of students excluded were actually very small and were not large enough to overpower the more stable tendencies of most states and territories. These analyses suggest that states probably used the same procedures for excluding students between the 1990 and 1992 TSAs.

To see if lower performing schools might have attempted to raise their NAEP scores by excluding more students, the investigator correlated average school scores with the total number of excluded students. A consistently negative correlation between average school scores and numbers of exclusions was found, suggesting a tendency for low-scoring schools to have larger numbers of special education and non-English-speaking students. It also seems to argue against the notion that schools or states were excluding large numbers of potentially low-scoring IEP or LEP students to artificially raise their scores. Despite the findings, low-scoring schools may have excluded more students to raise their NAEP scores but still earned low mean scores relative to other schools.

In comparing the scores of IEP and LEP students and non-IEP and non-LEP students, the investigator found that average performance for IEP and LEP students was very similar at both fourth and eighth grades. Distributions for IEP, LEP, and non-IEP and non-LEP low-scoring students were also very similar at both grades, indicating that the performance of IEP and LEP students is quite comparable to the performance of low-scoring students. Extended-response items, which require students to respond to open-ended questions by writing both an answer and a justification for it, were analyzed, and the most obvious finding was that few students in either the IEP-LEP group or the low-scoring non-IEP and non-LEP group performed very well on the items. While it was expected that differences in performance of 1 or 2 percentage points would be found, the results indicated that IEP and LEP students and a comparable group of low-scoring students performed equally badly on this aspect of the assessment.

If too many students were excluded from the assessment, it was hypothesized, those who were included would be able to complete it with few problems; thus completion rates were examined for IEP and LEP students who participated in the TSA and compared to those of low-scoring non-IEP and non-LEP students. It was expected that completion rates would be very high because observers of the TSA administration had noted that most students had no trouble finishing their booklets. Completion rates calculated across all booklets were quite similar for the IEP and LEP students and the comparison group, although in both grades 4 and 8 the rates were slightly higher for the IEP and LEP groups. Completion rates were also calculated on four selected blocks of items to examine student performance on blocks containing different types of items. Two blocks had mainly multiple-choice items; the other two were predominantly short-answer constructed response. (No extended-response items were included in these blocks.) It was found that on right-or-wrong items, a higher proportion of IEP and LEP students got items wrong and that a much higher

proportion of low-scoring students skipped items at the end of the block—they quit or gave up earlier. IEP and LEP students tended to attempt more items but also to get more items wrong. This suggests that IEP and LEP students are capable of taking the assessment since they attempted more items than did students in the comparison group.

An analysis was conducted to determine whether schools were excluding *all* identified IEP and LEP students from the assessment. Fifty percent of schools in the fourth-grade mathematics TSA and 28 percent of schools in the eighth-grade mathematics TSA excluded all identified IEP and LEP students from their assessments. However, a comparison of the average scores for these schools with the scores of other schools that included IEP and LEP students in their assessments showed very little overall difference in average school performance.

To get an indication of the comparability of special student populations between states, the variability of IEP and LEP student performance was examined across states in the TSA. At grade 4, mean mathematics scores for IEP students ranged from 161 (in Guam) to 212 (in Connecticut); the mean scores for LEP students ranged from 166 (in Rhode Island) to 235 (in New Hampshire). (Because there are fewer LEP students, standard errors of these means are high and means should be regarded with caution.) At grade 8, the differences were even larger: the mean scores for IEP students ranged from 183 (in Guam) to 252 (in North Dakota); for LEP students, from 184 (in Guam) to 285 (in North Carolina).

Type of community is a variable used in reporting NAEP data that is a rough measure of socioeconomic status. Mean performance of both IEP and LEP and non-IEP and non-LEP students tended to be highest among advantaged urban students and lowest among the disadvantaged urban. Within the IEP or LEP student category, it was found that IEP and LEP students in advantaged urban areas had higher average scores than non-IEP and non-LEP students in disadvantaged urban community schools. It is clear that student performance is affected by the community in which the student attends school. It is likely that the same types of results would be obtained if data were analyzed in terms of the type of community variable *and* race-ethnicity, but further investigation is necessary to understand the relationship between those variables and performance on the TSA.

Conclusions

The investigator concludes that her evidence argues both for and against the notion that the criteria for excluding students from the TSA may have been differentially applied in certain cases. As indicated above, the results support a conclusion that the exclusion criteria were probably applied consistently across the two TSA assessment cycles and that states were not manipulating the exclusion process to inflate their scores. On the other hand, some of the results hinted at the possibility that school personnel did not always strictly adhere to the exclusion criteria, especially the instruction to include "doubtful" students in the assessment. Participating IEP and LEP students completed most of the items in their booklets, suggesting that more IEP and LEP students could have participated in the TSA. Furthermore, comparisons of scores for IEP and LEP students with those of non-IEP and non-LEP students showed that although, as a group, the IEP and LEP students scored lower on the TSA, there was a

large amount of overlap in the two groups' score distributions, probably indicating that most of the IEP and LEP students taking the TSA were able to participate meaningfully. It may also mean that even more IEP and LEP students could have been assessed. Finally, many schools excluded *all* identified IEP and LEP students from participation, despite the fact that four-page questionnaires must be filled out for each excluded student. This should have served as a deterrent to categorical exclusions of IEP and LEP students. The categorical exclusion practice should be closely monitored in future assessments.

On the basis of these findings, four recommendations are made:

- Training for the assessment administrators should emphasize the importance of strict adherence to the exclusion guidelines for the assessment.
- An examination of how judgments are made to exclude students should be a part of future assessments. In particular, a process analysis of exclusions should be conducted in several different states, to help understand sources of variation in exclusions at the state and district levels.
- Future reporting of state-by-state results should include proportions of students excluded from each state's assessment, to reflect differences in exclusion patterns across states.
- Research efforts should be made to develop assessment tasks and modifications that can increase the participation of IEP and LEP students. Assessment modifications, accommodations, or alternatives could increase the partial participation of students with disabilities, with a goal of full inclusion.

State and Local Costs of the 1992 Trial State Assessment

Catherine O'Donnell
Jay Chambers
Dey Ehrlich
American Institutes for Research

In 1992 the National Assessment of Educational Progress (NAEP) implemented a Trial State Assessment (TSA) that included two subjects and two grade levels: fourth-grade mathematics, fourth-grade reading, and eighth-grade mathematics. Each of these three assessments involved collecting data on over 100,000 students, in over 4,000 schools, in 41 states, the District of Columbia, and 2 territories. The costs were shared by states, local school districts, and the federal government. The federal government paid for the

development of the assessment instruments, analyses and reporting of the data, and the training and monitoring of local test administrators. The states and local districts implemented the student sampling plan, administered the instruments, and collected questionnaires from teachers, counselors, and principals.

This study was undertaken to determine the costs incurred at the state and local levels in implementing the 1992 TSA. Because of the variety of budget and accounting procedures across states, the most meaningful cost estimates are those derived from information about personnel time expended and resources used. By comparing personnel efforts, participating states can learn how their implementation related to the average, without the confusion of differences in cost of living and salaries that exist from state to state.

State- and Local-Level TSA Activities

At the state level, a state coordinator worked with the prime contractor for carrying out NAEP and with the subcontractor for data collection, and had responsibilities in four major categories:

- **Meetings**—attended briefings on the TSA, meetings to discuss coordination procedures, and training for assessment administrators.
- **Recruitment**—secured the cooperation of public schools to participate in the TSA, working first with districts and then with assessment administrators to secure cooperation and participation.
- **Sampling**—reviewed the lists of schools selected in the subcontractor's sample selection; identified nonparticipating schools; and organized lists of fourth- and eighth-grade students in participating schools.
- **Coordination**—mailed the various support materials needed for the assessment (e.g., training materials, administration schedule forms, principal questionnaires); received and reviewed principal questionnaires and lists of eligible students.

At the local level, an assessment administrator had responsibilities in three major categories:

- **Preparing for the assessment**—arranged assessment sessions; attended a training session; identified students for exclusion; selected a supplemental sample of students; notified teachers and students of the assessment; preassigned and coded booklets; distributed and collected the assessment questionnaires; and scheduled makeups.

- **Conducting the assessment and makeup test administrations**—prepared the room, booklets, and other assessment materials; administered the assessment; and accounted for all materials.

- **Concluding the assessment**—completed the administration schedule; reviewed and returned materials; discussed the session with NAEP staff; and filed the administration schedule and roster of questionnaires.

Other school staff completed the excluded student questionnaire, teacher questionnaire, and school characteristics and policy questionnaire.

Methods

Letters were sent to the state coordinators in the 41 states, the District of Columbia, and 2 territories participating in the 1992 TSA, and the coordinators in 32 states and 2 territories agreed to participate. All 34 were sent a state coordinator survey questionnaire developed for this study. Within each state, a random sample of 24 sites (12 for fourth grade; 12 for eighth grade) was selected from the NAEP TSA sample to receive local assessment administrator survey questionnaires, with a total random sample of 805 schools selected. Only 18 sites were selected in each of the territories, 9 per grade.

To obtain the most accurate estimates of time spent on TSA implementation, the survey questionnaires were designed to be completed during the time that state coordinators and assessment administrators were involved in TSA activities. Participants were asked to record the time spent on the various activities and to return the questionnaires as soon as all of the activities were completed. Thus questionnaires had to be distributed before the assessments were administered, and strict time schedules had to be followed. For 65 schools, accurate information for mailing was not available within the required time frame, so these sites were not included in the final sample. In all, 740 schools received the assessment administrator questionnaire.

Both the state coordinator and assessment administrator survey questionnaires were designed to record the number of hours spent by each person on each activity. The state coordinator questionnaire also recorded each person's salary or hourly wage and the total number of hours worked per year. Travel and other expenses were recorded separately. Benefit rates were requested in terms of a percentage of salary or a dollar amount. Salary and benefit information for the assessment administrator questionnaire was collected from a third source and used to estimate hourly wage and benefit rates for each personnel category included on the questionnaire. An estimated average hourly compensation rate for personnel by job title was estimated for each state, based on state-by-state salary and benefit rate information. The assessment administrator questionnaire also requested the number of sessions held in the school by subject and grade.

All (100 percent) of the 34 state coordinator survey questionnaires and 609 (82.3 percent) of the 740 assessment administrator survey questionnaires were completed

and included in the database for analysis. Statistical imputation, through AIR's PROC IMPUTE (a hot-deck procedure) was used to assign values for missing data.

Results

The cost-study database included cost and time estimates for TSA activities from state coordinators for the total number of sessions held in each state, and estimates from the sample of local-level assessment administrators for the number of sessions held in each school sampled. As a result, population estimates for state-level costs were computed directly from the information provided on state coordinator survey questionnaires. Population estimates of total local-level costs were derived from sample estimates of local cost per session, then multiplied by the total number of sessions in each state. The unit of analysis for cost estimates was an assessment session, with state-level costs distributed over all the sessions held in a particular state. Costs for all three assessments were combined to obtain an estimate of the total cost, to each state and overall, of participation in the 1992 TSA.

Data were reported for a variety of factors and activities, by state. Overall, total costs were about $230,000 for full participation in the 1992 TSA (i.e., fourth-grade mathematics and reading and eighth-grade mathematics), with considerable variation from one state to another. The lowest cost was $81,175 (in Utah); the highest, $372,038 (in California). The cost per session was about $731, with 93 percent spent at the local level and 7 percent spent at the state level. In two states (California and New York), per-session costs were two standard deviations higher than the overall average.

When state- and local-level costs were combined, the fourth-grade assessment cost a state approximately $65,000; the eighth-grade assessment cost approximately $85,000. This difference resulted from the fact that in 95 percent of the fourth grades sampled, local-level costs for the fourth-grade assessment were allocated over two sessions within a school, while the local-level costs for the eighth-grade assessment were generally allocated over only one session. The average overall cost per session for the eighth-grade assessment ($806) was almost $300 higher than that of either of the fourth-grade assessments ($540), and it took, on average, 8 hours more of personnel time to conduct an eighth-grade session than to conduct a fourth-grade session.

Analyses of state-level costs by activity indicate that preparing for the assessment took the most time, mainly because of the time devoted to training local assessment administrators. Similar analyses of state-level estimates indicate that two activities, coordination and recruiting, in roughly equal proportions, accounted for 68 percent of personnel hours at the state level. Estimates for coordination varied most across states. Typically, about 2 hours of state-level activity were required per session: 40 minutes for coordination, 45 minutes for recruitment, 20 minutes for meetings, and 15 minutes for sampling. Nonpersonnel costs accounted for 29 percent of total state-level costs; personnel costs accounted for the remainder.

At the local level, costs per session varied considerably, depending on the assessment administrator's job title. It was most expensive to have the assistant superintendent fulfill this role; more often a teacher, principal, or instructional support staff member

was used. Cost estimates for eighth-grade assessments were generally higher than those for fourth-grade assessments because most schools held only one eighth-grade session. Total local-level costs were thus included in the cost for only one session.

Local-level time estimates for completing assessment questionnaires indicated that on average, 2 hours 30 minutes were needed to complete all questionnaires for fourth-grade sessions, and 3 hours 42 minutes for eighth-grade sessions. Most of this time was devoted to the mathematics and/or reading teacher questionnaire. Analyses of costs by type of community (extreme rural, disadvantaged urban, advantaged urban, and other) indicated that costs were less per session in extreme rural areas and were highest in eighth grade in disadvantaged urban areas.

Conclusions and Recommendations

On the basis of their findings, the investigators offered two recommendations:

- In view of the cost savings attained in 1992 by allocating costs over two fourth-grade assessments (mathematics and reading), it is recommended that future assessments be conducted in multiple subjects within a grade level.

Estimates of the cost involved in hiring a contractor to conduct the TSA were provided by a representative from the NAEP data-collection subcontractor. These estimates suggest that hiring a contractor could potentially reduce costs, for example in reducing local assessment-administration costs and eliminating quality control monitoring.

- Because districts that participated in the 1990 and 1992 TSAs are eligible to be sampled for future TSAs, it is recommended that consideration be given to hiring a contractor to conduct the TSA. This would serve to minimize the burden on districts that are asked to participate in the TSA, as well as potentially reduce the total cost of TSA implementation.

The Content and Curricular Validity of the 1992 NAEP Reading Framework

Bertram C. Bruce
Jean Osborn
University of Illinois at Urbana-Champaign
Michelle Commeyras
University of Georgia

The investigators undertook a study to analyze the content and curricular validity of the framework used for the 1992 National Assessment of Educational Progress (NAEP)

Trial State Assessment (TSA) in reading. The study procedures and findings are summarized below.

The Reading Framework for the 1992 NAEP

The Reading Framework for the 1992 NAEP lists the characteristics that identify good readers: good readers (a) exhibit positive reading habits and value reading; (b) read with enough fluency so that they can focus on the meaning of what they read; (c) use what they already know to understand the text they are reading; (d) extend, elaborate, and critically judge the meaning of the text; (e) plan, manage, and check the progress of their reading and use effective strategies to aid their understanding; and (f) read a wide variety of texts and can read for different purposes. Three "reading situations" are to be included in the assessment: reading for literary experience (e.g., the reading of novels, short stories, poems, plays, and essays); reading to be informed (e.g., the reading of magazine and newspaper articles, textbooks, encyclopedias, special interest books, and catalogues); and reading to perform a task (e.g., the reading of bus and train schedules, directions for games, recipes, consumer warranties, and office memos). The framework also identifies four aspects of reading: forming an initial understanding, developing an interpretation, personal reflection and response, and demonstrating a critical stance.

Regarding the assessment items, the Reading Framework calls for the use of a combination of open-ended and multiple-choice items and passage selections that use "authentic" texts, like those found in real, everyday reading. Extended passages selected for inclusion are to have coherence, orderly structure, and sufficient content to serve as the basis for items that can lead to meaningful student performance. Teacher evaluation rather than conventional readability formulas are to be emphasized in establishing the difficulty level of passages.

Analysis of the Consensus Process

Data were gathered from three sources: document reviews, interviews, and a survey.

Documents. Project staff examined the Planning Committee's letters, memos, the minutes of their meetings, and, a major source, the Committee's *Consensus Report*.[1]

Interviews. Personal and telephone interviews were conducted with members of the original Planning Committee; a group of reading educators and administrators not directly associated with the development of the 1992 NAEP reading assessment; and several members of the National Assessment Governing Board. Interview questions solicited views on the consensus process used to develop the framework, opinions

[1] Council of Chief State School Officers, *Report of the Consensus Process: 1992 NAEP Reading Assessment National Consensus Planning Project* (Washington, D.C.: Author, 1990).

about the framework's strengths (or weaknesses), and its adequacy in representing current theories and practices in reading education. Respondents were asked their opinion of the relative emphasis on open-ended responses and whether the 1992 NAEP provides a good model for reading instruction.

Survey. A survey designed to solicit reactions to the framework's content was mailed to a random sample of 250 participants in the National Reading Conference, a random sample of 250 presidents of International Reading Association (IRA) councils; all 50 chief state school officers; all 50 state directors of Chapter I reading programs; all 50 state reading specialists; everyone who had responded in writing to earlier drafts of the framework; and leaders in the IRA and the National Council of Teachers of English (NCTE). For the 700 surveys sent out, 288 responses were received. Time restrictions precluded a followup mailing. An additional 1,000 questionnaires were sent to classroom teachers who were members of IRA, and 339 teachers responded. The combined results from the two surveys (627 responses) were reported.

Results—The Adequacy of the Consensus Process

A major goal of the consensus process was to induce the interest and cooperation of key figures in reading and major professional education organizations. The project staff and the Steering and Planning Committees invited broad-based participation and communicated the results of committee deliberations. Interviews conducted with consensus-process participants indicated that all involved were working to develop a framework that would deserve broad support, especially given the time constraints.

Individual members of the Planning Committee made positive comments about the development process. Nonetheless, five areas of concern about the consensus process emerged from an analysis of the documents and the interviews: (1) the time for the development of consensus, particularly given the known divisions within the reading community; (2) the involvement of major professional organizations, with the decision of the IRA to disassociate itself from the process identified as particularly troublesome; (3) state and local reactions affecting support, including the state-reporting aspect of NAEP, the national discussion about lifting the prohibition against reporting district-level NAEP data, and the lack of systematic analysis of state and local curricula and assessment policies and objectives for the Reading Framework; (4) coordination of the various aspects of the total development process, due mainly to severe time constraints but problematic in some cases; and (5) the quality of public documents, such as the evidence of haste in the Reading Framework (not surprising given time limitations and the number of people involved); the clarification needed for concepts associated with constructing, extending, and elaborating meaning; and the lack of more specific headings and subheadings.

Results—The Consensus of the Field

Most (78 percent) survey respondents were teachers at the elementary, secondary, college, or university level. More than 72 percent of the respondents said they were

somewhat or very familiar with NAEP. A large majority (93 percent or more) agreed with each of the characteristics of good readers presented in the framework, and more than 95 percent found the framework's definitions of reading either acceptable or very acceptable.

When asked their opinions about the proportion of items allocated at each grade level to each type of reading situation, most (60 percent) of the respondents disagreed with the allocation but varied widely in their suggestions for percentages of items for each grade level and reading situation. Regarding the three reading situations specified by the framework, more than 88 percent of the respondents said it was either essential or very important to assess these situations. About 85 percent of the respondents said no other aspects of reading needed to be represented in the assessment.

The respondents were very supportive of the inclusion of open-ended items: only 4 percent objected to them. Considerable support (67 percent) was shown for the decision to use only authentic passages (that is, full-length texts of the type that students are likely to encounter in everyday reading), and a majority of respondents indicated that the framework had successfully met each guideline set out by the Steering Committee.

Expert Judgments of Key Features

Two groups of reading professionals were convened to analyze the framework. The first was a colloquium of approximately 50 people, mostly faculty and graduate students at the University of Illinois College of Education; the second was a panel that included three professors of education, one member of a state board of education, one public school administrator, the executive directors of IRA and NCTE, and one education consultant. Nearly all approved of the attempt to align the assessment with the process of reading, the use of open-ended questions, the use of authentic texts, and student choice.

Most of the experts seemed pleasantly surprised by the general approach taken in the framework and its reflection in the assessment. They approved of the view of "reading as a constructive, dynamic process, not just the assembly of a set of subskills." The experts also agreed with the emphasis on open-ended responses but recognized the challenge of scoring them. Virtually all agreed that another positive feature was the use of naturally occurring, whole, "authentic" passages, and they approved of giving choices to 12th-grade students as to what they read for the assessment.

The consensus from the interviews, the panel, and the colloquia was that alignment with the reading process, open-ended responses, authentic texts, and student choice were evidence for the validity of the assessment. But some concerns also arose, as described below.

1. **Measuring reading through writing.** The colloquium participants agreed that the 1992 NAEP in reading is unusual in incorporating writing as a response mode for reading, a significant advance in large-scale assessment. However, they were

concerned that writing ability would be confounded with reading ability and suggested that other response modes be considered for future assessments.

2. **Reading situations.** The assessment's notion of situation is problematic, implying a one-to-one correspondence between text genres and reading situations. This contradicts the notion of "reading situation" as it is usually employed, a situation defined by the reader's purpose, the social context, the task, and only in part by the text. The implied one-to-one correspondence can have negative pedagogical consequences and implies that children should not be encouraged to learn from, or be informed by, a literary text. This inference contravenes recent recommendations about using reading in the content areas.

3. **Aspects of reading.** The assessment attempts to assess the ability of readers to deal with four aspects of reading, but instead of seeing stance as a response that varies widely among individuals, including skilled readers, or even during the reading of a passage, the assessment treats stance as imposed by the task. Thus, the combination of a passage and a question causes the reader to adopt a particular relationship to the passage. This apparently subtle shift in definition to meet assessment needs leads to a very different model, one that was problematic for the participants in the colloquia, for the members of the expert panel, and for the project staff.

Summary and Recommendations

Overall, the investigators concluded that the consensus process for the 1992 Reading Framework was adequate, if somewhat rushed, and the resultant document is reasonably responsive to most of the current theories and practices in reading. The Reading Framework does about as well as any in representing the views of an enormously complex and divided field. The following recommendations are offered to improve the next cycle of development:

- A strong center of responsibility should be created to oversee all aspects of work involved in developing, administering, scoring, and reporting a National Assessment of Educational Progress in reading. This center of responsibility would permit a closer coordination among the institutions and groups involved in the various aspects of developing and implementing NAEP.

- Participation of experts from the field of reading should extend beyond the planning stages.

- More time should be allocated for the planning process; otherwise, it is not possible to build a wide consensus.
- Documents such as the framework that are produced for public dissemination should be clarified and made more consistent.

Evaluation of the 1992 Reading Framework for the National Assessment of Educational Progress

Julia H. Mitchell
American Institutes for Research

A Reading Framework was developed for the National Assessment of Educational Progress (NAEP) through a national consensus process, establishing the theoretical underpinnings and specifications for constructing the 1992 reading assessment. The National Academy of Education commissioned two previous studies of the content validity of the reading assessment: (1) a 1991 analysis of the framework and the consensus process used to develop it and (2) a 1993 examination of the "fit" of the NAEP reading items to the framework.[1] To capture current thinking and perspectives about how well the Reading Framework reflects consensus among members of the reading community, a survey of reading experts was undertaken in October 1993.

Method

From names submitted by P. David Pearson of the University of Illinois and Alan E. Farstrup of the International Reading Association, reading experts from around the country were contacted, sent copies of the Reading Framework, and asked to answer four questions about the framework, either in telephone interviews or in written responses. Of 23 reading experts contacted, 21 participated: 8 in interviews and 13 through written responses. The four questions and a summary of the responses are presented below.

1. ***What is your view of the 1992 NAEP framework as a basis for a national assessment? What are its strengths? its weaknesses?***

Nearly all (19) of the 21 reading experts felt that the 1992 framework was a major improvement over earlier NAEP approaches to assessing reading. Of the 21

[1] See Bruce, Osborn, and Commeyras, "The Content and Curricular Validity of the 1992 NAEP Reading Framework," and Pearson and DeStefano, "Content Validation of the 1992 NAEP in Reading: Classifying Items According to the Reading Framework," both summarized in this volume.

respondents, 3 gave only a general endorsement of the framework, while the other 18 were more specific about characteristics that they considered to be particular strengths. Those most commonly mentioned were:

- the use of different types of texts and reading purposes; taking into account the interactions between reader, text, and context; and giving a holistic view of a person as a reader in a variety of authentic reading situations and contexts (cited by 14 respondents)

- the assessment strategies that move away from multiple-choice to open-ended responses and the use of longer and authentic passages (cited by 9 respondents)

- the focus on outcomes (embedding evaluation within a performance-based approach) and away from reading as a set of discrete skills (cited by 7 respondents)

The overall response to the framework was clearly positive.

Along with this overall positive response, however, a number of the respondents expressed major concerns about NAEP's capacity to implement the assessment promised by the framework. In large part, these concerns were based on the quality of the sample passages, items, and scoring rubrics published with the framework,[2] although there were also some more general concerns about the limitations of NAEP's paper-and-pencil, on-demand assessment format. For example, some respondents noted that the assessment placed little emphasis on the communicative context in which reading takes place and required students to write answers without the opportunity for reflection and discussion encouraged by current reading and writing theory.

On the other hand, a few respondents felt that the framework had departed too far from traditional reading assessment. They felt, for example, that the assessment could underestimate the reading abilities of students who lack the requisite writing skills to respond successfully to the open-ended items which comprise a large portion of the assessment.

2. Do you think that the framework adequately represents current theory and practices in the field of reading education? How? How not?

The responses to questions 1 and 2 were often intertwined because factors cited as strengths in response to question 1 were often those that the respondent believed were best aligned with current theory and practice, while the factors cited as weaknesses were those the respondent considered least aligned. In general, respondents were supportive of the changes from previous NAEP frameworks, for example, in assessing with a variety of response modes and using different text types and reading situations. Areas in which respondents felt that the framework does *not* reflect current theory and practice included the absence of naturalistic tasks to provide context, inadequate attention to important aspects of integrated language instruction

[2] The sample items and scoring rubrics published in the framework were prototypes developed before the rest of the item pool. Plans call for them to be replaced with more carefully crafted materials when the framework is reissued for the 1994 assessment cycle.

(e.g., the role of discussion in shaping the understanding of and response to text), and the use of scoring criteria that imply a preferred interpretation of the way the reader is supposed to respond to a task and develop *the* meaning.

3. ***Overall, what grade would you give to the framework, given the purpose for which it is intended?*** (Use "A" to represent the best possible grade and "F" to represent the worst.)

Responses to this question provided another indication that most of the reading experts approved of the framework: 2 gave it an "A"; 16 gave it a B+, B, or B-; and 3 gave it a C+, C, or C-. No one gave it a grade below C-.

4. Periodically, NAEP undertakes major revisions of the assessment in specific content areas. When this is done, a new framework is developed and a new trend line for scores is started. The 1992 Reading Framework represented one such "fresh start." ***Given the changes you see going on in the reading field, how many years do you think it will be before NAEP will have to create a new, revised framework? Why?***

Responses to question 4 ranged from "immediately, already out of date" (3 responses) to "at least 20 years" (1 response), with 15 of the 21 falling in ranges of the 2-5 years (8 responses) and 5-10 years (7 responses). Two could not say. Some respondents qualified their answers by suggesting that the process should be one of review, not necessarily revision, and some asserted that the real issue is refining the *assessment* so that it is more representative of the framework content.

Summary

A majority of the 21 reading experts surveyed were positive about the NAEP Reading Framework, seeing it as a major improvement over earlier NAEP frameworks and a good effort at achieving consensus. They approved of its constructivist definition of reading which takes into account the effects of reader, text, and context on the construction of meaning; the attention to different types of text and reading purposes; the movement away from the multiple-choice response mode; the use of longer and authentic passages; the focus on processes and outcomes and away from skills; and the emphasis on critical thinking, reflection, and analysis. Overall, the 21 respondents gave it a solid "B" in fulfilling its intended purposes. About three-fourths (76 percent) of the respondents predicted that the framework could be retained for 2 years or considerably longer.

The major concerns expressed were related to the implementation of the framework. Several respondents felt that the conceptualization of reading articulated in the framework could be adequately assessed in a paper-and-pencil, on-demand assessment like NAEP. Some experts were concerned that, while the framework proposes to assess reading in a variety of situations and acknowledges that meaning is influenced by context, the student being assessed is reading only for the purpose of completing the assessment. Other experts noted that although the integration of reading and writing through the use of open-ended items was a major step forward

for the framework, the assessment still falls short of measuring the social aspects of reading and the integration of reading (and writing) with speaking and listening.

Finally, respondents expressed some concern about the sample items and scoring rubrics included in the framework. These samples were completed before the development of the main item pool and will be replaced when the framework is reissued for the 1994 assessment cycle.

The Content and Curricular Validity of the 1992 NAEP in Mathematics

Edward A. Silver and Patricia Ann Kenney
Learning Research and Development Center
University of Pittsburgh

The investigators continued a study of mathematics items that they began with the 1990 National Assessment of Educational Progress (NAEP) in mathematics.[1] For this second study, they analyzed the 1992 NAEP framework and items used in the 1992 grade-4 and grade-8 Trial State Assessment (TSA) in mathematics. One of the study goals was to determine the relationship between the NAEP framework and test content and the goals espoused in the National Council of Teachers of Mathematics (NCTM) Standards. Another was to study important changes made at the item level, particularly the increase in the number of items and the addition of extended open-ended items and manipulative items.

Analysis of the NAEP Framework and Items

For both grades 4 and 8, the 1992 NAEP mathematics items were organized into 13 blocks; each item had two classifications, one for content area (numbers and operations; measurement; geometry; data analysis, statistics, and probability; algebra and functions) and one for mathematical ability (conceptual understanding, procedural knowledge, and problem solving). The investigators compared the intended percentages in the 1992 Mathematics Framework to the actual percentages of items in the NAEP classifications with respect to ability and content. An expert panel of seven mathematics education professionals representing different constituencies and perspectives examined the items with respect to the NAEP ability and content categories and to selected themes from the NCTM Standards. They also reviewed the three special item types (extended open-ended, manipulative, calculator block) and identified exemplary and flawed items.

[1] E.A. Silver, P.A. Kenney, and L. Salmon-Cox, "The Content and Curricular Validity of the 1990 NAEP Mathematics Items: A Retrospective Analysis," in *Assessing Student Achievement in the States: Background Studies* (Stanford, CA: The National Academy of Education, 1992).

Findings of the Study

The 1992 NAEP TSA in mathematics, at both grades 4 and 8, was an improvement over the 1990 NAEP TSA in terms of content validity and was more closely aligned with the NCTM Standards than was the 1990 NAEP. In particular, the actual distribution of NAEP items within the ability categories for both the grade-4 and grade-8 assessments deviated from the distribution specified by the NAEP framework with a trend toward including more items classified as problem solving and conceptual understanding and fewer classified as procedural knowledge. These findings suggest that the Item Development Committee was striving to create a better match between test content and current thinking about important aspects of mathematical performance.

Analysis of other aspects of the 1992 NAEP TSA test content also provided evidence of greater alignment with current thinking about mathematical performance. At both grades, the number of items increased, providing greater coverage of content areas, and the percentage of open-ended items increased. The two new item types (manipulative and extended open-ended) introduced in the 1992 mathematics TSA were found by content experts to be representative of the kinds of items that should be included on NAEP mathematics assessments.

The finding that almost 60 percent of items from the grade-4 TSA were judged to match at least one major theme from the NCTM Standards indicated an improvement in alignment of NAEP test content with the standards since only about 50 percent of the 1990 grade-8 TSA items had been judged by a similar panel to match any of the major themes of the NCTM Standards. In general, the link between item types and standards themes was judged by the panel of content experts to be particularly strong in the case of open-ended and extended open-ended items. Because these items require students to communicate mathematically, this finding is consistent with observations and arguments made in other published sources.

Concerns about the Content Validity

Although the findings of this study provided some positive indicators of the content validity of the 1992 NAEP TSA in mathematics, they also revealed a variety of concerns, of which some were similar to those noted for the 1990 TSA and others were new to the 1992 TSA. Among the lingering concerns was the existence of ambiguity in the definitions of NAEP ability categories. Two independent groups of content experts, working with different sets of items (i.e., the 1990 grade-8 TSA item pool and the 1992 grade-4 TSA item pool) at different times, experienced similar difficulties in classifying items with respect to the NAEP mathematics ability categories, clear evidence of an inadequacy in the descriptive information accompanying the Mathematics Framework that guided development of the 1990 and 1992 NAEP assessments. This ambiguity casts doubt on the usefulness of attempts to draw fine distinctions between conceptual and procedural knowledge on the basis of student performance and on the validity of ability-oriented interpretations that might be made on the basis of student performance on NAEP.

Another concern relates to the identified deficiencies in the 1992 grade-4 TSA item pool. The finding that almost 60 percent of the items were judged to match at least one major theme from the NCTM Standards has an alternative interpretation: over 40 percent were judged to not match any of the four major themes. Also, despite the presence of innovative item types and a new set of calculator-block items, the grade-4 expert panel identified only about 10 percent of the grade-4 items as exemplary. The expert panel judged the calculator-active items to be especially unimaginative, a criticism quite similar to one made by the panel reviewing the 1990 grade-8 TSA calculator-block items. Innovative item formats (manipulative and extended open ended) were perceived as a welcome addition to the 1992 NAEP TSA in mathematics, but brought a new set of concerns, including difficulties associated with task wording and the timing and placement of these items. The use of a focused holistic scoring model raised concerns about reliability and validity issues, for example, discrimination between adjacent score levels.

Recommendations for Future NAEP Mathematics Assessments

The National Assessment Governing Board approved a framework for the 1994 NAEP mathematics assessment that includes important changes, such as using a model consisting primarily of the five major content strands used in the 1990 and 1992 assessments. The investigators reached the tentative conclusion that the framework is in an evolutionary mode, that it more closely reflects the vision for school mathematics presented in the NCTM Standards, and that many of the changes are compatible with recommendations made in the investigators' evaluation of the content and curricular validity of the 1990 NAEP TSA. It will be important for the NAEP mathematics assessment to continue to move toward greater alignment with the NCTM Standards.

Three recommendations for future NAEP mathematics assessments are made:

- The amount of time for development, review, and critique in all aspects of the NAEP mathematics assessment must be increased. Evidence suggests that the current 2-year cycle authorized for mathematics creates increasing overlap between activities associated with contiguous assessments. The decision to increase the frequency of administration of NAEP should be carefully considered in terms of content quality and validity, implications for developing and field testing new item types, and allocation of time for developing and reviewing the frameworks, items, and scoring guides.

- A coherent, sustained, reflective process of monitoring and shaping future NAEP mathematics assessments needs to be put into place immediately, and the process must be deeply informed by those most knowledgeable about school mathematics and education. In their report on the 1990 NAEP TSA, the investigators called for the establishment of an oversight

committee to monitor all aspects of NAEP mathematics achievement-levels creation, framework development, item writing, test construction, and reporting of results, and they echo that recommendation, envisioning a group of classroom teachers, curriculum experts, members of the lay public with subject-matter expertise, and researchers. This would give the participation of content experts greater focus and provide coherence to the evolution of NAEP over time.

- Greater convergence is needed between the ideas and language used to frame the design of the NAEP mathematics assessment and those used to report student performance results. The decision to use the 1992 NAEP mathematics achievement levels as the primary reporting method for the 1992 mathematics results makes this recommendation especially important. The lack of fidelity between the test content and the achievement-level descriptions (as noted in the investigators' separate study on mathematics achievement levels) raises serious questions about the validity of drawing inferences about what students should be able to do from test performance. It is strongly recommended that a more planful approach be put into place in which achievement-level descriptions are carefully constructed so that they become an integral part of the Mathematics Framework, serve as guidelines for item development, and are used as the basis for appropriate reporting categories for NAEP results.

Content Validation of the 1992 NAEP in Reading: Classifying Items According to the Reading Framework

David Pearson
Lizanne DeStefano
University of Illinois at Urbana-Champaign

The Reading Framework was designed to present a rationale and to provide guidelines for the scope and format of the 1992 National Assessment of Educational Progress (NAEP) for reading. As a document, it was intended to reflect consensus in the field as well as sound, contemporary research about reading. As a content framework, it was intended to guide item development for the 1992 NAEP reading assessment and to serve as the basis of the achievement-level descriptors.

Developed over a 12-month period by a number of committees comprised largely of professional educators, the framework proposed some major changes in the content and format of NAEP reading, including:

- a definition of *a good reader* based upon a constructivist view of learning

- use of authentic texts
- classification of three reading situations
- assessment of cognitive aspects of reading
- an increased emphasis on open-ended questions

Each of these changes presented new challenges for the developers of NAEP and the field as a whole. More important, they created a new set of assumptions that would affect how the 1992 NAEP in reading would be constructed, how achievement levels would be set, and how results would be interpreted.

To assess the content validity of the assessment, this study convened a panel of experts to judge the extent to which items on the 1992 NAEP in reading match the content framework proposed in the framework. Data collected via this process were intended to answer the following research question:

> To what extent do items on the 1992 NAEP in reading represent the domain of reading as described in the Reading Framework?

In addition, categorizations assigned by these panelists were compared with those assigned by staff members of the NAEP contractor to determine level of agreement on an item-by-item basis.

Method

A panel of 12 experts was convened for a 1-day session to review all of the items, passages, and scoring criteria for the 1992 NAEP in reading for each grade level and to categorize each item (considered in the context of its corresponding passage and scoring criteria) in terms of a three-by-four matrix. The matrix was constructed by pairing each of the three reading situations (reading for literary experience, reading for information, and reading to perform a task) with each of the four reading aspects (initial understanding, developing interpretation, personal reflection and response, and demonstrating a critical stance) as described in the Reading Framework.

The 12 panelists (4 at each grade level) were all educators with extensive experience teaching or supervising reading at specific grade levels; all were familiar with state and national assessment initiatives, including NAEP; and each was an active member of at least one professional organization in reading or language arts, such as the International Reading Association, the National Reading Council, or the National Council of Teachers of English. All were familiar, at least in a general sense, with the Reading Framework and the NAEP in reading.

Before the session, panelists were sent copies of released NAEP passages and items and a copy of the Reading Framework. At the session, the framework was reviewed, and panelists discussed the three-by-four reading-aspect-by-situation matrix, developed common ways of defining each situation and aspect, and identified factors that seemed to differentiate among cells in the matrix. Panelists were given strategies

for using characteristics of the passage, the format and content of the item or prompt, and the criteria used to evaluate students' responses in arriving at a classification for an item. Then the group as a whole categorized items for a sample passage and item set using an analysis form designed for the study.

Panelists then broke into grade-level groups and were given copies of every passage, all items, and the scoring guides for that grade level. Each participant categorized all of the items for a passage according to the three-by-four reading-aspect-by-situation matrix. Then in grade-level groups, panelists compared categorizations, discussed any areas of disagreement, and arrived at consensus as to the classification of the items for that passage. The same series of activities was carried out for each passage at that grade level. After all the items were categorized, panelists were asked to make a summary sheet by compiling the consensus categorizations across all passages.

Facilitators were available to provide assistance with interpretation and application of the descriptors, and they observed within-group deliberations. At the end of the full-day session, a full-group debriefing was held, and each grade-level group summarized its perceptions of the extent to which the items of the 1992 NAEP in reading corresponded to the Reading Framework.

Findings

In several instances, panelists had difficulty assigning a single situation-by-aspect category to an item. That difficulty reflects a common criticism of the framework, that is, that it implies mutual exclusivity across each combination of reading aspect and reading situation. For example, in the case of reading situation, the framework implies that students read fictional stories "for literary experience" only and never "to be informed." Similarly, the rigidity of the classification scheme suggests that an item that requires students to demonstrate a critical stance cannot also tap personal response or reflection. When panelists felt that an item involved more than one situation or aspect, the investigators allowed them to place it in two cells but with one indicated as the primary classification. Only primary classifications were used for computations.

The relationship of items on the 1992 NAEP in reading to the Reading Framework was considered in three ways:

- by examining the distribution of the items across reading situations
- by looking at the allocation of assessment time across reading aspect
- by comparing the panelists' situation-by-aspect categorization of each item against the classification assigned by the NAEP contractor

Major findings of the study were as follows:

- ***Distribution across reading situation.*** The panelists' categorization of the items on the 1992 NAEP in reading closely

approximated the distribution targeted in the item specifications and in the Reading Framework.

- ***Distribution across reading aspect.*** Based on the panelists' categorizations, substantially less assessment time was devoted to personal reflection and response or to demonstrating a critical stance than was called for by the Reading Framework. Correspondingly more time was spent on items that required the student to form an initial understanding or develop an interpretation.

- ***Consistency of categorization.*** High levels of agreement were found among panelists and between panelists and the contractor on the categorization of reading situation. Agreement was harder to achieve for reading aspect, and panelists tended especially to disagree with the official NAEP categorizations for personal-response and critical-stance items. This was attributed to ambiguity in the definitions of these two categories and to inconsistencies between these definitions and the rubrics used to score items.

Recommendations for Future Assessments

The study produced three recommendations for consideration in the design of future assessments:

- The panelists felt that the distribution of items by reading situation recommended by the consensus panel, and specified by the framework, should be reconsidered. There was strong feeling that items assessing the ability to read to perform a task should be included at the 4th grade as well as at the 8th and 12th grades.

- Based on the panelists' judgment, very few items on the assessment provided an adequate opportunity for students to demonstrate their capacity for personal response or their ability to take a critical stance. In many instances, the concern lay with the criteria that were used for scoring the constructed-response items. In the future, it is recommended that these scoring procedures be carefully considered to ensure that they are consistent with the definitions of reading situations and aspects described in the framework.

- Some general ambiguities in the definitions of reading situations and aspects were also noted. In the future, these definitions should be clarified and clearly illustrated using example items from the assessment and the accompanying rubrics.

Impact of the 1990 NAEP Trial State Assessment Program: A Followup Study

Frances B. Stancavage
American Institutes for Research
Edward D. Roeber
Council of Chief State School Officers
George W. Bohrnstedt
American Institutes for Research

This study continues an evaluation of the impact of the 1990 National Assessment of Educational Progress (NAEP) Trial State Assessment (TSA). An earlier report examined prerelease preparations and the immediate postreport activities carried out by the states.[1] This report looks primarily at the longer term changes (positive or negative) that occurred as a result of the TSA program or its results.

Methodology

Data were gathered from three sources:

1. Periodic **telephone surveys** were conducted with samples of five groups of end-users most concerned with the direction of mathematics education and assessment: 28 national educators (e.g., members of the NAEP Planning and Steering Committees, current presidents of national education organizations represented on those committees), 15 national policymakers (House and Senate committee staff), 98 state assessment directors and mathematics specialists, 62 state educators and policymakers, and 105 local educators (district superintendents, principals, curriculum and assessment specialists). Three rounds of interviews were conducted: 2 months before the first release of NAEP TSA results (on June 6, 1991) and 2 months and 6 months after this initial release. During the intervening time, the data were rereleased (on September 30, 1991) by the National Assessment Governing Board (NAGB) and the National Education Goals Panel (NEGP), using the newly developed achievement levels which set a standard for basic, proficient, and advanced performance.

2. During the fall of 1992, **case studies** were developed for a subset of nine states that participated in the 1990 TSA: California, Colorado, Connecticut, Delaware, Florida, Indiana, Louisiana, Nebraska, and North Carolina. These included all of the states

[1] F.B. Stancavage, E.D. Roeber, and G.W. Bohrnstedt, "A Study of the Impact of Reporting the Results of the 1990 Trial State Assessment: First Report," in *Assessing Student Achievement in the States: Background Studies* (Stanford, CA: The National Academy of Education, 1992).

that reported during the telephone surveys that NAEP had had a significant impact on their instruction and/or assessment. Extended, semistructured interviews were carried out with each state's assessment director and mathematics specialist, and case reports, designed to place the release and its impact in the broader context of the state's unique needs and circumstances, were developed. After review and approval by the respondents, the case studies were included in the investigators' report.

3. **News releases, newspaper reports, and other artifacts** from the reporting process were examined for completeness and accuracy and for analysis of the types of information reported. News stories about NAEP that appeared from June through December 1991 in the 50 major national newspapers were obtained from a national news clipping service. National press conferences held in conjunction with the data releases were attended by staff, and the press materials distributed there were reviewed.

Research Findings

The report is organized around nine questions, which are presented below, followed by summaries of the relevant findings.

1. **How and to what extent did the states disseminate the information released by NAGB and NEGP on September 30, 1991?** Dissemination by state departments of education was much more limited for the September 30 release than for the June 6 release of results. Fewer than 40 percent of the participating states issued press releases (as compared to 97 percent in June), and the numbers preparing speeches or articles, holding press conferences, or providing personal briefings were even more limited. Respondents attributed the decreased attention to the fact that the September release only repeated information already reported in June and indicated that their actions did not necessarily reflect a preference for one reporting format over the other.

2. **How did the news media handle the reporting of the September 30 release? Was media coverage more or less extensive than it was for the initial release in June? more or less accurate?** The September 30 release received dramatically less press coverage than the June 6 release. More than half were straightforward news stories focusing on the reports released and including comments from various education leaders. Unlike the articles that followed the June 6 data release, most of those appearing in October 1991 did not devote space to state comparisons. Survey respondents corroborated the impression formed from reading the clippings that press coverage of the September release was quite limited, but their consensus of opinion was that the coverage was fair and factual.

3. **Were the traditional anchor levels (used in the June 6 reports) or the new achievement levels (used in the September 30 reports) preferred?** More than 50 percent of the state department of education employees most responsible for disseminating the NAEP results felt that achievement levels were more relevant for communicating the state of education and more likely to have an impact on education policy. Less than 25 percent had a clear preference for anchor levels in these contexts, and most of the remainder felt that the choice depended upon the situation in which the results were to be used. National educators and policymakers tended to lean toward the anchor levels, but 40 percent or more responded with a "don't know" to questions comparing anchor levels and achievement levels.

4. **As time progressed, were the 1990 state NAEP results still being discussed or used? Had they given rise to meaningful discussions among educators and policymakers?** The interview respondents in two-thirds of the states reported that activities precipitated by the June release of information were still taking place in July and August 1991, and a high proportion of the respondents also reported that the release of TSA results had provoked discussions of mathematics performance among key constituencies. In January-February 1992, respondents reported that the TSA results had continued to stimulate discussion among key user groups, with as many or more respondents reporting discussions in January-February as reported them in July-August.

5. **To what extent has information about the NAEP TSA filtered down to the local district and school level?** Awareness of NAEP appears to have increased somewhat. Compared to the local sample interviewed just before the June release (in April 1991), 67 percent of a new randomly drawn local sample recognized NAEP by name and an additional 17 percent were familiar with the concept. However, only one-third of those surveyed remembered seeing any of the reports of the TSA or hearing anything about the results. Respondents who had seen the results were overwhelmingly (90 percent) of the opinion that the results were presented in a way that allowed for understanding and appropriate interpretation; only a few were aware of the results actually being used by their districts to improve mathematics instruction or mathematics assessment.

6. **What evidence is there that the state NAEP program is influencing mathematics instruction?** The survey results show that changes in mathematics curriculum, instruction, and teacher training were pervasive among states, particularly with new curricular materials, new teacher education programs or requirements, and alignment with the standards developed by the National Council of Teachers of Mathematics (NCTM). Among those making changes, more than one-fourth (27 percent) reported being influenced by NAEP/TSA as of July-August 1991 and a significantly higher percentage (42 percent) reported this in

January-February 1992. While the influence of the TSA thus appears to be reasonably widespread, both the surveys and case studies made it clear that NAEP carries weight not in isolation but because it articulates well with other contemporary influences in mathematics education. Some specific uses and influences attributed to NAEP/TSA included tipping the balance in favor of calculators (in the classroom and on assessments) and using sample NAEP items as models for their own assessments and for purposes of teacher in-service training.

7. **What evidence is there that the state NAEP program is influencing mathematics assessment?** A similar pattern was observed with regard to the influence of the NAEP/TSA on mathematics assessment: the influence of the TSA was more likely to be seen as additive rather than unique. Specifically, respondents indicated that the TSA had hastened the passage of accountability legislation and the adoption of new or revised state assessments. They also reported that NAEP items had influenced the design of state assessments.

8. **What specific contextual factors influenced the impact of the TSA results in individual states?** The impact of the TSA on instruction or assessment at the classroom level appears to be mediated through the state departments of education and/or legislatively mandated changes in state assessments or teacher requirements. The absence of local-level data in NAEP limits the saliency of the results for local districts. Moreover, even if such data were available, it is not clear that they would compete successfully with data from state assessments that are more directly tied to state curriculum and local accountability. Trade-offs and linkages between NAEP and state assessments are still evolving, and clearly the course of this evolution will greatly affect the ways in which state NAEP data are used in the future. At present it appears that the state-level impact of NAEP is inversely related to the investment that the state has in its own assessment system.

9. **Overall, is the state NAEP program a useful exercise?** More than 2 years after the first TSA, the NAEP state assessments appear to have been accepted as an established player in the assessment arena. While little evidence of penetration below the level of the state legislature and state departments of education was found, the majority of stakeholders at these levels reported a positive impact on education in their states and an expectation that impact would increase once the results of the 1992 TSA were released. Education and policy observers at the national level, however, were somewhat less sanguine, with over a quarter rating the 1990 NAEP mathematics assessment as not very successful in providing information that will help educators and policymakers improve U.S. education. The majority of respondents, both state and national, expected the impact of the TSA to increase after the release of the 1992 data.

Issues in the Development of Spanish-Language Versions of the National Assessment of Educational Progress

Walter G. Secada
University of Wisconsin at Madison

Debates about the content and organization of the National Assessment of Educational Progress (NAEP), its role in education reform, the students included and excluded by its design, and its possible use in measuring progress toward national education standards afford a rare opportunity for thinking about the development of Spanish-language assessments to measure the achievement of Spanish-speaking students. The purpose of this study is to consider issues and options surrounding the possible development of Spanish-language versions of NAEP for use in the United States.

The Larger Social Context

The number of students in U.S. schools who speak a language other than English at home is rapidly increasing. In 1990, the total number of U.S. residents 5 years and older who speak a language other than English at home was 31,845,000—14 percent of the residents 5 years and over and an increase of almost 38 percent over the total in 1980.[1] For these individuals, by far the most common language is Spanish, accounting for more than half (54.4 percent) of this group and more than 10 times the number who speak the next-most-common language. Between 1980 and 1990, the number of residents 5 years and older who speak Spanish at home increased by 50.1 percent, a trend that continues.

Possible Sources of Bias in an All-English-Language NAEP

One of the primary reasons proffered for developing NAEP in other languages is to reduce the bias, or unfairness, that may result from using an all-English-language NAEP to test language-minority students. This potential source of bias involves two opposite effects. The first effect is that excluding all students with limited English proficiency (LEP) may result in systematic overestimates of student achievement. The second, opposing, effect is that student achievement may be underestimated in cases where students who are literate in a non-English language are tested in English.

LEP student exclusions. NAEP is designed to exclude students on the basis of language proficiency. However, the criteria could result in differential student selection because they specify exclusion that is mediated, at least in part, by the availability of bilingual education. Because more schools have English-as-a-second-language programs than have bilingual education programs, it seems likely that

[1] U.S. Census data.

schools with bilingual education programs would exclude more LEP children than would other schools. Local school policies and practices and definitions of language proficiency also vary, increasing the possible bias of results.

Excluding students on the basis of their English proficiency is likely to result in overestimates of Hispanic achievement on NAEP, with biases that are greater in reading than in mathematics.[2] As NAEP moves toward the use of open-ended and/or authentic tasks for assessing mathematics performance, the increased language load is likely to result in greater bias of mathematics results. Variability in procedures for excluding students makes it difficult to pinpoint the size of the overestimate, but as the Hispanic student population grows, the biases are likely to increase.

Optimal performance in another language. Another source of bias is the possibility that some students are tested in their weaker language and would perform better if tested in their stronger language. Hence the English-language NAEP underestimates the performance of these students and of the racial and ethnic groups to which they belong. The extent of this potential bias is not known.

Most Spanish-dominant LEP students are clustered in the lower elementary grades. Because the modal experience for Spanish-speaking LEP students is to be enrolled in all-English-speaking programs, the amount of their exposure to the Spanish language drops very quickly after primary school. When LEP students are taught via Spanish, they tend to receive minimal literacy development and academic instruction, with primary focus on ideas that they did not grasp in English. Many students maintain their Spanish competence, but this is generally because of experiences and influences outside the classroom.

Are there enough students whose Spanish-language literacy is dominant over their English-language literacy so that underestimates of their performance on an all-English NAEP would significantly bias the results? This question has not been answered. Since there are countervailing sources of language bias in NAEP, the actual effects on estimates of any group's performance may be negligible. However, the credibility given to NAEP data goes beyond technical questions, and the potential bias becomes a fundamental threat to NAEP's usefulness in informing policy, particularly in states with large numbers of Hispanic students.

Current Uses of Assessments in Spanish

The investigator conducted a survey to determine how Spanish-language assessments are being used in areas with high concentrations of Spanish-speaking students. Telephone interviews were conducted with individuals in state departments of education in six states that together enroll 75 percent of the nation's Hispanic students and in seven urban districts that enroll large numbers of Hispanic students.

All six states surveyed allow for the exclusion of LEP students from their mandated testing programs; all six recommend Spanish-language testing to identify special needs among Spanish-speaking, bilingual children. Only New York and Illinois use Spanish-

[2] See B.D. Spencer, "A Study of Eligibility Exclusions and Sampling: 1992 Trial State Assessment," also summarized in this volume.

language testing as part of their regular testing program. In New York, the mandated ninth-grade Regents Examination is available in Spanish, as are competency tests in several subjects and primary-grade pupil and program evaluation tests. In Illinois, Spanish-language tests are used to evaluate bilingual programs. Other states prefer to exclude LEP students from their assessment programs for reasons that include the costs of developing and administering tests in other languages; reluctance to subject LEP students to the difficulty of taking English-language assessments; and, absent data, the ability to avoid blame for the educational status of non-English-speaking students.

At the district level, six of the seven districts surveyed use Spanish-language tests to assess students excluded from English-language tests in the ongoing testing programs. Other reported uses are in identifying special needs among Spanish-speaking, bilingual children; determining students' proficiency levels in Spanish at enrollment and at regular intervals thereafter; and meeting evaluation requirements of federally funded bilingual education programs. All reports of Spanish-language testing were for students in the elementary grades.

Reasons for Developing Spanish-Language Versions

Spanish is the logical language to use for a non-English NAEP, and precedents for Spanish-language assessments are found in large urban districts and some states. In Puerto Rico, Spanish-language versions of NAEP are being developed in mathematics and science. Another reason for a Spanish-language NAEP is the issue of equity. Some have argued that, without safeguards to ensure educational equity, the educational status of many ethnic minorities and children from lower socioeconomic backgrounds will suffer as a result of assessment-reform efforts. These individuals point to the history of testing in the United States, where many different sorts of tests have been used to misclassify students, to support prejudices, and to constrain educational opportunities. A Spanish-language assessment could be one way to promote an equity agenda: it is simply fairer to test individuals in their strongest languages.

Another reason is to promote curriculum and program development. One could argue that an overreliance on English-language tests (for purposes of entry, exit, and placement of students and program evaluation) has resulted in systematic constraints on educational programs for language-minority students, with emphasis on low-level English-language skills and the development of remedial programs that focus on those skills.

Other reasons include increasing the motivation of states and urban districts to participate in NAEP, since Spanish-language versions should help fill some important data gaps and report on student proficiency with greater accuracy and inclusiveness; and developing new technologies for test equating and studying dual-language competence.

Reasons for Not Developing Spanish-Language Versions of NAEP

One reason offered for not developing a Spanish-language version of NAEP is that student diversity can be addressed in other ways: through assessment tasks and items that are readily understandable by many student populations; choices among tasks that are structurally equivalent but varied in context; and the inclusion of items on student language proficiency and opportunity to learn to assist in interpreting assessment results.

Another reason is the lack of resources and the need to focus the limited resources available on curriculum, teaching, and programs. Comparability issues are also raised: ensuring comparable scoring of English and Spanish constructed-response items; equating scores; accounting for cultural differences; and developing a Spanish-language assessment that would be appropriate for students with different Spanish dialects.

Recommendations

The investigator finds the arguments supporting the development of Spanish-language versions of NAEP to be more compelling and cites the need to inventory and marshal all of the U.S. human resources, including its linguistic resources and its Spanish-speaking students' knowledge. This potential for developing economic policy based on better knowledge of our foreign-language resources tips the balance for him. He hastens to caution that not everyone agrees with him in favoring the development of a Spanish-language assessment. Many Hispanics expressed reluctance or opposition because of the scarcity of resources and, for some, a keenly felt distrust of any assessment program and the uses to which it might be put.

The investigator offers two recommendations.

- That a pilot study be conducted to determine the effects on patterns of exclusion and achievement levels of Spanish-language versions of NAEP: the credibility of NAEP results and their usefulness in informing policy are at stake. He proposes a small-scale study in reading and mathematics for elementary school students. For the proposed pilot, the investigator outlines procedures that have been used successfully in his research. The study currently underway in Puerto Rico, to develop Spanish-language versions of NAEP for mathematics and science, uses these procedures and could also provide useful information for the proposed pilot study.

- That ongoing NAEP policies and practices be developed that are consonant with student diversity and equity. Equity must be viewed in the gestalt, not simply as an issue of controlling for bias in test results or of developing Spanish-language versions of a test. At core, the issue is one of the overrepresentation of students from specific demographic groups in educational

categories that make it easy to stereotype students and to constrain their educational opportunities. Equity must be manifested in the commitment of policymakers and assessment practices that ensure fairness. Ongoing concern and sensitivity to equity issues must be developed and demonstrated.

The Judged Congruence Between Various State Assessment Tests in Mathematics and the 1990 National Assessment of Educational Progress Item Pool for Grade-8 Mathematics

Lloyd Bond
Richard M. Jaeger
Center for Educational Research and Evaluation
University of North Carolina at Greensboro

In 1990 for the first time since its inception more than 20 years ago, the National Assessment of Educational Progress (NAEP) provided, for 37 states, the District of Columbia, and 2 U.S. territories, state-level results in eighth-grade mathematics comparable to results for the nation. Several states have indicated an interest in linking state NAEP results to their own assessments for comparative purposes. Another study directed by Robert Linn and sponsored by the Technical Review Panel for NAEP will result in empirical calibrations of state mathematics assessments for eighth graders in several states to the NAEP grade-8 mathematics item pool and scale.[1] The present study examines the content congruence of the same set of state mathematics tests to the 1990 NAEP eighth-grade item pool. By examining the congruence for a number of states, this study yields results that, when combined with the results of the Linn study, provide the possibility of establishing a threshold index of congruence that supports acceptably stable equating of state assessments to NAEP.

Procedures

Three state mathematics tests were used:

- the California Achievement Tests (CAT), Form 3, Level 18
- the Metropolitan Achievement Tests (MAT), Advanced 1, Form 1
- the Stanford Achievement Test (SAT), Eighth Edition, Advanced 1, Form J

(The CAT and SAT were also used in the Linn study.)

[1] R.L. Linn and V.L. Kiplinger, "Linking Statewide Tests to the National Assessment of Educational Progress: Stability of Results" (Boulder: University of Colorado, Center for Research on Evaluation, Standards, and Student Testing, 1993).

Following a small pilot study, the investigators assembled a panel of five mathematics education experts for a 2-day session to judge the content congruence of these three tests to the NAEP item pool for eighth-grade mathematics. The NAEP content framework for mathematics consists of five subject-matter categories (numbers and operations; measurement; geometry; data analysis, statistics, and probability; and algebra and functions), crossed with three mathematical ability categories (conceptual understanding, procedural knowledge, and problem solving).

Working independently, members of the expert panel were required first to place the items on each test into one of the NAEP subject-matter categories or into an "unclassifiable" category, and then to place each item assigned to a subject-matter category into one of the three mathematical ability categories or into an "unclassifiable" category. Following their initial categorizations of items, panel members discussed items for which there was not unanimous or near-unanimous agreement and then made a second set of item-placement judgments, again working independently.

The proportion of items placed in each of the 15 categories of the NAEP content framework for eighth-grade mathematics was computed for each state mathematics assessment test and for each panelist and was reported by panelist and by average across panelists. Corresponding proportions of NAEP items placed in each content-framework category were also reported.

Content Congruence of the State Assessment Tests to NAEP

The results indicated three general patterns that appeared to characterize the experts' item-classification judgments for all three state assessment tests:

1. The expert panel classified a significantly larger proportion of items into the cell of the NAEP content framework defined by the numbers-and-operations subject matter and the procedural-knowledge mathematical ability category. Forty percent of the items on the MAT and 51 percent of the items on both the SAT and the CAT were assigned to this category, compared to only 14 percent of the NAEP grade-8 mathematics items in this category.

2. The panelists tended to classify relatively more items into the numbers and operations/conceptual understanding category, compared to the proportion of NAEP items in this category. Far less notable than the first pattern, this pattern was most pronounced for the MAT, where 28 percent of the items were placed in the numbers and operations/conceptual understanding category (after discussion), compared to only 13 percent of the NAEP items in the category.

3. In part as a consequence of the first two patterns, the panelists assigned low proportions of items to the other categories in the NAEP content framework. Thus, although 25 percent of the NAEP items purport to reflect problem-solving ability, only 11 percent of the CAT items, 5 percent of the MAT items, and 10 percent of

the SAT items were judged by panelists to measure this ability. Measurement was the only category where the proportions of assigned items on the state assessment tests were relatively close to the corresponding NAEP proportions.

Indices of Content Congruence

For this study, four statistical indices were developed to summarize the content congruence between the state assessment tests and the NAEP item pool: two based on the squared difference between the proportion of NAEP items in the NAEP categories and the proportion of state test items in the NAEP categories, and two based upon the corresponding absolute differences in these proportions. All four indices have the desirable property that larger and larger deviations from perfect agreement with the distribution of items in the NAEP item pool are reflected in smaller and smaller values for each index.

Overall, the investigators are persuaded that the two indices that are based on squared proportional differences have the most desirable statistical properties: (1) they are the most general, in the sense that they can be applied to *all* possible classifications of state assessment items in the NAEP content framework, including situations where some or all items are not classifiable within the content framework, and (2) the function relating the values of these two indices to various degrees of difference between the distributions of state assessment test items and NAEP grade-8 mathematics items, over the cells of the NAEP content framework, is a monotonically decreasing one. The two indexes, called the J-Index and the JH-Index, are likely to yield similar values in most cases, and the choice is a matter of personal preference for squared, or absolute-value, distance metrics.

With a potential range of zero (indicating no congruence) to 1.0 (indicating perfect congruence), the J-index values for the CAT, MAT, and SAT were, respectively, 0.615, 0.676, and 0.616. These values suggest substantial differences between the distributions of state test items and NAEP items, across the cells of the NAEP content framework.

Interjudge Agreement

Substantial agreement was found across all pairs of judges as to the placement of state assessment test items within the NAEP content framework. In general, members of the expert panel demonstrated a high level of agreement on the classification of state assessment-test items within the subject-matter categories of the NAEP content framework, but much lower levels of agreement on the mathematical abilities measured by the state assessment-test items. The J-Index of congruence was applied to the item classifications of randomly selected pairs of judgment-panel members, yielding an average of 0.84.

Conclusions and Recommendations

At the outset of this study, it was hoped that the three state achievement tests examined would differ markedly in their content congruence with the 1990 NAEP mathematics item pool. However, the MAT, SAT, and CAT proved to be quite similar to one another in their content distributions and to differ from the NAEP mathematics item pool almost to the same degree. The investigators would not, therefore, expect any of the three tests to be more readily equatable to the NAEP mathematics item pool than are the others.

The results of this study suggest that states currently using the standardized achievement tests examined in this study would best refrain from attempts to equate them to the NAEP 1990 mathematics item pool. Investment in a state assessment measure that more clearly mirrors the content distribution of the NAEP item pool is recommended for states that want to link their assessment tests to NAEP. For example, Kentucky administers a state assessment test that was modeled after the NAEP mathematics item pool.

The investigators strongly recommend a replication of this study and that of Linn, using the 1992 and, when available, 1994 statewide assessment data from Kentucky and any other states that can provide consistent student achievement-test data for both school years. Such a replication would provide an essential test of the equatability of a state assessment test to NAEP under ideal conditions and might yield threshold values of indices of congruence that would guide states in their future equating decisions. The investigators also recommend a replication of their study using 1992 and 1994 NAEP and statewide reading-assessment data for fourth graders. Since the interpretive benefits of reliable linkages between state assessment tests and NAEP item pools would likely be substantial, future investment in this line of inquiry is strongly recommended.

A Study of the Administration of the 1992 National Assessment of Educational Progress Trial State Assessment

Elizabeth Hartka
Donald H. McLaughlin
American Institutes for Research

As part of the evaluation of the Trial State Assessment (TSA) of the National Assessment of Educational Progress (NAEP), a study of the administration of the 1990 TSA was carried out by the American Institutes for Research (AIR). The purpose was to assess state-by-state NAEP for feasibility in terms of cost, feasibility of implementation by local staff, and validity of the results. The report's conclusion was that the 1990 TSA was carefully planned and implemented and that training was adequate for the assessment; however, recommendations for future refinements of assessment procedures were made. For the 1992 TSA, a new level (grade 4) and a

new content area (reading) were added, and more states participated. Because of these changes in size and scope, AIR undertook a second evaluation of the TSA, described below.

Classroom Observations

To witness the administration of the TSA, AIR staff visited schools in northern California, Colorado, Massachusetts, New Hampshire, and Rhode Island. Although the visits were unsystematic and the schools chosen were not randomly sampled, a good sense of how well the administration actually took place from the school's point of view could be discerned. It was particularly instructive to see how the students reacted to the assessment.

Prior to making classroom observations, AIR staff attended training sessions for quality control monitors and assessment administrators (AAs). The school visits were coordinated with state supervisors for Westat (a NAEP subcontractor), and sites were balanced between monitored and unmonitored sessions and within reasonably close driving distance of one another. Some balance between high socioeconomic status (SES) and low SES schools was achieved in New England. Observers arrived before the start of the session and watched the assessments as unobtrusively as possible, taking extensive written notes and using copies of quality control forms developed by Westat for recording information about the administration.

Results from the observations were grouped around three research questions:

- How closely did the AAs follow the instructions? In most sessions observed, AAs made sincere efforts to follow the instructions, and errors seemed to be errors of omission. A strong tendency to embellish the script or instructions for fourth graders was noted, but the changes appeared to make the instructions more meaningful.

- How adequately did the AAs complete their tasks? Most AAs were found to be extremely diligent about following instructions; however, some tasks were overlooked several times.

- How valid was the assessment from an observer's point of view? Almost all observers noted irregularities with the paced-tape administration of the estimation items in the mathematics assessment.

Overall, observations of the administration of the 1992 TSA indicated that it went well. Although some problems were noted, they were minor. The mathematics assessment was more complicated to administer and took about 30 minutes longer than the reading assessment. This complexity arose from the need to distribute and redistribute calculators and to monitor the use of rulers, protractors, and manipulatives. Students were also observed to become restless toward the end of the assessment, particularly in fourth grade, and to have difficulty in providing background information.

A paper-and-pencil questionnaire, similar to one used in evaluating the 1990 TSA, was mailed to state testing directors in September 1992. Six weeks later, followup phone calls were made to those who had not yet responded. Written responses were received from all participating states and territories and all nonparticipating jurisdictions except one.

Patterns of results were very similar across the two assessment years. In most cases (66 percent), the 1992 respondents said that the final decision to participate was made by the state superintendent of education. The most frequently stated reason for participation (cited by 64 percent of respondents) was that states wanted to have a voice in shaping NAEP policy; second (cited by 61 percent) was to obtain state ranking information. No one reported that the administration of the assessment did not go well. The two most common problems reported were getting schools to participate and requiring too much staff time for administration, both cited by 48 percent of respondents. When asked to assign an overall grade to the assessment, 23 percent said "A"; 55 percent, "B"; 18 percent "C"; 2 percent D; 0 percent F; and 2 percent missing, all improvements from the 1990 percentages of 15, 55, 23, 2, 0, and 5, respectively, for 1990.

Regarding the conduct of the TSA, 59 percent of the state testing directors felt that more time was required than necessary, an increase from 43 percent in 1990, and 59 percent felt that more paperwork was involved than necessary, an increase from 57 percent in 1990. Solid support for the amount of training involved in the administration was expressed: 86 percent described the amount as about right, an increase from 70 percent in 1990.

When asked about policy issues, 68 percent of the state testing directors expressed interest in linking state NAEP to existing assessments in their states; 52 percent were interested in linking state NAEP to international assessments; and slightly under half (46 percent) favored a full blown state-by-state assessment in 1994/1996. Most states in 1992 reported that they planned to participate in the TSA in 1994, but their affirmative responses were not quite as definite as they had been in 1992. More states cited budgetary approval as a condition for participation (21 percent in 1992 versus 10 percent in 1990), and, in response to a new question about what would increase the likelihood of their participation in the next round, the most frequently cited condition was reimbursement to states for partial costs of the assessment paperwork and administration. The most preferred subjects and grade levels for assessment in 1994 were fourth-grade reading and mathematics, closely followed by eighth-grade reading and mathematics.

State testing directors in nonparticipating states and territories reported that the principal reasons their states declined to participate were costs, availability of funds, and increased burden on schools and staff. The same reasons were given in 1990, but in the more recent survey many additional reasons were listed. Costs were still important to 64 percent of the nonparticipants, but having a choice about subjects and grade levels in the assessment was chosen by about 90 percent of these respondents.

Three different approaches were used to address the validity of the results: (1) results of the 1992 TSA were compared to results from an equivalent sample of students from the 1992 national assessment; (2) data from *monitored* classrooms were compared with data from *unmonitored* classrooms; and (3) student performance in monitored classrooms where specific irregularities in session administration occurred was compared with performance in classrooms where no irregularities occurred. Standardized logit percent scores (measures of performance of fourth-grade reading and fourth- and eighth-grade mathematics assessments) were the dependent variable in the analyses. Details about the sessions were obtained from Westat forms and enabled investigators to examine whether 18 factors had effects on student performance. Analyses involved variance component calculations and the Student's t tests of significance; results focused on three research questions:

1. **Were there differences between the TSA and the national assessment in terms of student performance?** In 1990, it was found that students in the TSA obtained slightly but reliably higher scores in eighth-grade mathematics. In 1992, however, there were no significant differences between student performance in the TSA and in the national comparison sample.

2. **Were there differences between monitored and unmonitored classrooms?** Most differences were very small; the exception was for fourth-grade reading, where the difference was statistically significant. However, the difference was of only slightly larger magnitude than the other *non*significant differences, and it seemed reasonable to conclude that performance in monitored and unmonitored sessions was about the same.

   A related question was whether there were differences between the results for schools in the original sample and those for substitute schools. Across all grades and subject areas, the mean performance levels for the substitute schools in 1992 were slightly higher than the means for the original schools, as they were in 1990. It was concluded that for the eighth grade, the original and substitute schools produced similar results as described by mean levels of student performance and in the amount of variation of students. Significant differences in mean levels of student performance were found for fourth grade, for reasons that were not clear.

3. **Did student performance vary with respect to specific irregularities in NAEP administrative procedures?** Taken together, the findings indicated that differences in mean-level performance across all 1992 grades and subjects in the following areas related to the session administration of the TSA: ability of the AA to work independently, student cooperation, student attitude toward the assessment, and the status of the AA.

The following recommendations are made:

- Although most state testing directors reported that the 1992 TSA went well, the percentage reporting problems with getting schools to participate doubled between 1990 and 1992, as did the percentage who thought that the assessment took too much staff time. Because of the increased burden, particularly for small states, a study should be conducted of the feasibility of using half-samples and allowing states to participate in the subject and grade of their choice.

- Monitoring of the sessions provided valuable information and probably avoided many problems. It should be continued at some level; however, the results of the analyses suggest that monitoring could probably be reduced by a small amount.

- Training of the assessment administrators should continue to emphasize the value of uniformity of administration procedures.

- More care should be taken in developing administration procedures for younger NAEP students.

- Better background measures, particularly for fourth-grade students, should be collected for the TSA.

Notes

Alexander, L. and James, H.T. *The Nation's Report Card: Improving the Assessment of Student Achievement*. Washington, D.C.: The National Academy of Education, 1987.

Beaton, A.E. "Methodological Issues in Reporting NAEP Results at District and School Levels." Paper prepared for the National Assessment Governing Board, October 2, 1992.

Bond, L. and Jaeger, R.M. "The Judged Congruence Between Various State Assessment Tests in Mathematics and the 1990 National Assessment of Educational Progress," in *The Trial State Assessment: Prospects and Realities: Background Studies*. Stanford, CA: The National Academy of Education, forthcoming.

Bruce, B.C., Osborn, J., and Commeyras, M. "The Content and Curricular Validity of the 1992 National Assessment of Educational Progress Reading Framework," in *The Trial State Assessment: Prospects and Realities: Background Studies*. Stanford, CA: The National Academy of Education, forthcoming.

Council of Chief State School Officers. *Report of the Consensus Process: 1992 NAEP Reading Assessment National Consensus Planning Project*. Washington, D.C.: Author, 1990.

Council of Chief State School Officers. *Report of the Consensus Planning Process: 1990 NAEP Mathematics Assessment*. Washington, D.C.: Author, 1988.

Dossey, J.A., Mullis, I.V.S., and Jones, C.O. *Can Students Do Mathematical Problem Solving? Results from Constructed Response Questions in NAEP's 1992 Mathematics Assessment*. Washington, D.C.: National Center for Education Statistics, August 1993.

Educational Testing Service, handout on distribution of assessment time by item format in 1994 and 1992, NAEP Network meeting, June 18, 1993.

Educational Testing Service. *The NAEP Guide: A Description of the Content and Methods of the 1990 and 1992 Assessments*. Princeton, NJ: Author, 1990.

Goals 3/4 Technical Advisory Group meeting, Council of Chief State School Officers. Washington, D.C., April 29, 1993.

Hartka, E. and McLaughlin, D. "A Study of the Administration of the 1992 National Assessment of Educational Progress Trial State Assessment," in *The Trial State Assessment: Prospects and Realities: Background Studies*. Stanford, CA: The National Academy of Education, forthcoming.

Hartka, E. "A Study of Students Excluded from the 1992 National Assessment of Educational Progress Trial State Assessment," in *The Trial State Assessment: Prospects and Realities: Background Studies*. Stanford, CA: The National Academy of Education, forthcoming.

International Reading Association. *Assessment Resolutions*. Newark, DE: Author, 1991.

Johnson, E.G., Mazzeo, J., and Kline, D.L. *Technical Report of the NAEP 1992 Trial State Assessment Program in Mathematics*. Washington, D.C.: National Center for Education Statistics, 1993.

Kaufmann, J.M. "How We Might Achieve the Radical Reform of Special Education," *Exceptional Children*, 60, 1993.

Kirsch, I.S., Jungeblut, A., Jenkins, L., and Kolstad, A. *Adult Literacy in America*. Princeton, NJ: Educational Testing Service, September 1993.

Koretz, D., Skewes-Cox, T., and Burstein, L. "Omitted and Not-Reached Items in Mathematics in the 1990 National Assessment of Educational Progress." University of California at Los Angeles: Center for Research on Evaluation, Standards, and Student Testing, September 1992.

Koretz, D.M., Linn, R.L., Dunbar, S.B., and Shepard, L.A. "The Effects of High-Stakes Testing on Achievement: Preliminary Findings About Generalization Across Tests." Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, April 1991.

Linn, R.L. "Linking Results of Distinct Assessments," *Applied Measurement in Education*, 6, 1993.

Linn, R.L. and Kiplinger, V.L. "Linking Statewide Tests to the National Assessment of Educational Progress: Stability of Results." Boulder, CO: Center for Research on Evaluation, Standards, and Student Testing, 1993.

Linn, R.L., Koretz, D.M., Baker, E.L., and Burstein, L. *The Validity and Credibility of the Achievement Levels for the 1990 National Assessment of Educational Progress in Mathematics*. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing, January 1991.

McLaughlin, D.H., Stancavage, F.B., Chambers, J.G., Hartka, E., and Ercikan, K. "Evaluation of the Implementation of the 1990 Trial State Mathematics Assessment," in *Assessing Student Achievement in the States: Background Studies*. Stanford, CA: The National Academy of Education, 1992.

Miller, L.S., Fredisdorf, M., and Humphrey, D.C. *Student Mobility and School Reform*. New York, NY: Council for Aid to Education, 1992.

Mislevy, R.J. *Linking Educational Assessments: Concepts, Issues, Methods, and Prospects*. Princeton, NJ: Educational Testing Service, December 1992.

Mullis, I.V.S., Campbell, J.R., and Farstrup, A.E. *NAEP 1992 Reading Report Card for the Nation and the States*. Washington, D.C.: National Center for Education Statistics, September 1993.

Mullis, I.V.S., Dossey, J.A., Owen, E.H., and Phillips, G.W. *NAEP 1992 Mathematics Report Card for the Nation and the States*. Washington, D.C.: National Center for Education Statistics, April 1993.

Mullis, I.V.S., Dossey, J.A., Foertsch, M.A., Jones, L.R., and Gentile, C.A. *Trends in Academic Progress: Achievement of U.S. Students in Science, 1969-70 to 1990; Mathematics, 1973 to 1990; Reading, 1971 to 1990; Writing, 1984 to 1990*. Washington, D.C.: National Center for Education Statistics, November 1991.

Mullis, I.V.S., Dossey, J.A., Owen, E.H., and Phillips, G.W. *The STATE of Mathematics Achievement*. Washington, D.C.: National Center for Education Statistics, June 1991.

NAEP Reading Consensus Project. *Reading Framework for the 1992 National Assessment of Educational Progress*. Washington, D.C.: National Assessment Governing Board, 1992.

National Academy of Education. *Setting Performance Standards for Student Achievement*. Stanford, CA: Author, 1993.

National Academy of Education. *Assessing Student Achievement in the States*. Stanford, CA: Author, 1992.

National Assessment of Educational Progress. *1994 National Assessment of Educational Progress Mathematics Assessment: Framework and Specifications*. Princeton, NJ: Educational Testing Service, March 1992, unpublished report.

National Assessment Governing Board. *Positions on the Future of the National Assessment of Educational Progress, 1993*. Washington, D.C.: March 1993.

National Assessment Governing Board. *The Future of NAEP*. Washington, D.C.: Author, August 1992.

National Center for Education Statistics. *NAEP 1992 Mathematics State Report for California*. Washington, D.C.: Author, April 1993.

National Center for Education Statistics. *1992 TSA: Summary of Participation Rates, Training, and Data Collection Activities*. Washington, D.C.: Author, July 1992.

National Center for Education Statistics. *Summary for All States, Training, and Data Collection Activities, 1992 Trial State Assessment*. Washington, D.C.: Author, May 1992, draft.

National Commission on Excellence in Education. *A Nation at Risk*. Washington, D.C.: U.S. Department of Education, 1983.

National Council on Disability. *Serving the Nation's Students with Disabilities: Progress and Prospects*. Washington, D.C.: Author, 1993.

National Council on Education Standards and Testing. *Raising Standards for American Education*. Washington, D.C.: Author, January 24, 1992.

National Council of Teachers of Mathematics. *Curriculum and Evaluation Standards for School Mathematics*. Reston, VA: Author, 1989.

National Education Goals Panel. *The National Education Goals Report: Building a Nation of Learners*. Washington, D.C.: Author, 1991.

O'Donnell, C., Chambers, J.G., and Ehrlich, D. "State and Local Costs of the 1992 Trial State Assessment," in *The Trial State Assessment: Prospects and Realities: Background Studies*. Stanford, CA: The National Academy of Education, forthcoming.

Office of Educational Research and Improvement. *Trends in Academic Progress*. Washington, D.C.: Author, 1991.

Pearson, D. and DeStefano, L. "Content Validation of the 1992 NAEP in Reading: Classifying Items According to the Reading Framework," in *The Trial State Assessment: Prospects and Realities: Background Studies*. Stanford, CA: The National Academy of Education, forthcoming.

Pearson, D. and DeStefano, L. "An Evaluation of the 1992 NAEP Reading Achievement Levels, Report Three: Comparison of Cutpoints for the 1992 NAEP Reading Achievement Levels with Those Set by Alternate Means," in *Setting Performance Standards for Student Achievement: Background Studies*. Stanford, CA: The National Academy of Education, 1993.

Phillips, G. et al. *Interpreting NAEP Scales*. Washington, D.C.: National Center for Education Statistics, 1993.

Public Law 100-297, Part C, Section 3403 6A: 1988.

Public Law 100-297, Part C, Section 3402: April 28, 1988.

Report on Educational Research. August 18, 1993.

Rivera, C. and Pennock-Roman, M. *Issues in Race/Ethnicity Identification Procedures in the National Assessment of Educational Progress, Part I: A Comparison of Observer Reports and Self-Identification*. Research Report No. RR-87-19. Princeton, NJ: Educational Testing Service, 1987.

Secada, W.G. "Issues in the Development of Spanish-Language Versions of the National Assessment of Educational Progress," in *The Trial State Assessment: Prospects and Realities: Background Studies*. Stanford, CA: The National Academy of Education, forthcoming.

Silver, E.A., and Kenney, P.A. "The Content and Curricular Validity of the 1992 NAEP TSA in Mathematics," in *The Trial State Assessment: Prospects and Realities: Background Studies*. Stanford, CA: The National Academy of Education, forthcoming.

Silver, E.A. and Kenney, P.A. "Expert Panel Review of the 1992 NAEP Mathematics Achievement Levels," in *Setting Performance Standards for Student Achievement: Background Studies*. Stanford, CA: The National Academy of Education, 1993.

Silver, E.A., Kenney, P.A., and Salmon-Cox, L. "The Content and Curricular Validity of the 1990 NAEP Mathematics Items: A Retrospective Analysis," in *Assessing Student Achievement in the States: Background Studies*. Stanford, CA: The National Academy of Education, 1992.

Spencer, B.D. "A Study of Eligibility Exclusions and Sampling: 1992 Trial State Assessment," in *The Trial State Assessment: Prospects and Realities: Background Studies*. Stanford, CA: The National Academy of Education, forthcoming.

Spencer, B.D. "Eligibility/Exclusion Issues in the 1990 Trial State Assessment," in *Assessing Student Achievement in the States: Background Studies*. Stanford, CA: The National Academy of Education, 1992.

Stancavage, F.B., Roeber, E.D., and Bohrnstedt, G.W. "Impact of the 1990 NAEP Trial State Assessment Program: A Followup Study," in *The Trial State Assessment: Prospects and Realities: Background Studies*. Stanford, CA: The National Academy of Education, forthcoming.

Stufflebeam, D., Jaeger, R.M., and Scriven, M. *Summative Evaluation of the National Assessment Governing Board's Inaugural 1990-91 Effort to Set Achievement Levels on the National Assessment of Educational Progress*. Washington, D.C.: National Assessment Governing Board, August 1991.

Truby, R. "The Future of NAEP as the Nation's Report Card." Washington, D.C.: National Assessment Governing Board, November 29, 1989.

U.S. Department of Commerce. "Number of Non-English Speaking Americans Up Sharply in the 1980s, Census Bureau Says," *U.S. Department of Commerce News*. Washington, D.C.: April 1993.

U.S. General Accounting Office. *Educational Achievement Standards: NAGB's Approach Yields Misleading Interpretations*. Report No. GAO/PEMD-93-12. Washington, D.C.: Government Printing Office, June 1993.

Westat/NAEP. *Manual for Assessment Administrators, Fourth Grade, 1992 Trial State Assessment*. Rockville, MD: Author, 1992.

Acronyms Appearing in this Report

AA	Assessment Administrator
AAAS	American Association for the Advancement of Science
ACT	American College Testing
AIR	American Institutes for Research
BIB	Balanced Incomplete Block
CAT	California Achievement Tests
CCD	Common Core of Data
CCSSO	Council of Chief State School Officers
ECS	Education Commission of the States
EIAC	Education Information Advising Committee
ETS	Educational Testing Service
GAO	General Accounting Office
IDEA	Individuals with Disabilities Education Act
IEP	Individualized Education Plan
IRA	International Reading Association
IRT	Item Response Theory
LEP	Limited English Proficiency
MAT	Metropolitan Achievement Tests
NAE	National Academy of Education
NAEP	National Assessment of Educational Progress
NAGB	National Assessment Governing Board
NCES	National Center for Education Statistics
NCEST	National Council for Education Standards and Testing
NCS	National Computer Systems
NCTE	National Council of Teachers of English
NCTM	National Council of Teachers of Mathematics
NEGP	National Education Goals Panel
NESIC	National Education Standards and Improvement Council
NSF	National Science Foundation
QED	Quality Education Data, Inc.
SAT	Stanford Achievement Test
SES	Socioeconomic Status
SREB	Southern Regional Education Board
TIMSS	Third International Mathematics and Science Study
TRP	Technical Review Panel
TSA	Trial State Assessment

BEST COPY AVAILABLE